The Dissertation Committee for Raghav Shroff
certifies that this is the approved version of the following dissertation:

Biology in the information age:
Computational methods to understand and engineer the central dogma

Committee:

Andrew Ellington, Supervisor

Bryan Davies

Ron Elber

Vishwanath Iyer

Edward Marcotte

Biology in the information age:
Computational methods to understand and engineer the central dogma

by

Raghav Shroff

Dissertation

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

The University of Texas at Austin
May 2020

To my family,

for their unwavering support

Acknowledgments

Many times throughout my graduate school journey I didn't think I would reach the finish line. I'd like to give some space to thank those who pushed me along the way.

First and foremost, thank you to my advisor Andy Ellington. The scientific ownership you afford your students turns this place into a contagious atmosphere filled with inspiring research. I am grateful you allowed me to figure out what my own interests are even through times when I was floundering and accepted my ways of thinking even when they ran contrary to your own. I doubt I could have succeeded in any other lab.

To Ross Thyer, thanks for being a mentor and the needed skeptic. I am in awe of your vast scientific knowledge and persistence in pursuing good science. I can confidently say you have made me a better scientist.

I believe innovative ideas are borne when you have a colleague willing to push your boundaries and challenge your creativity. For me, this was Austin Cole. Thanks for being a sounding board for new ideas and constantly inspiring me to think bigger. I look forward to see how far we can go.

To Ellington lab members past and present, I will cherish the inspiring and amusing times we've had together. Special thanks to Jared, Jimmy, Johnny, Andre, Danny, Shaunak, Liz, Simon, Alex, Stella, Barrett, and Mason.

Finally, to Aycan, you are my rock. Thanks for believing in me even when I don't.

Abstract

Biology in the information age:
Computational methods to understand and engineer the central dogma

Raghav Shroff, PhD

The University of Texas at Austin, 2020

Supervisor: Andrew Ellington

The rise of NGS, big data, and '-omics' has ushered biology into a new age, with the power to fundamentally change how research is approached. Rather than using a singular hypothesis, we can now incorporate more data-driven methods that drive new biological insights, explain emergent biological phenomena, and/or derive novel functionality. This thesis highlights the changing role of computation to both learn more about biological systems as well as leveraging data-intensive computational techniques to create new proteins and enzymes.

The ability for computational approaches to drive biological understanding is presented in three studies. First, the laboratory evolution of DNA polymerases, the workhorses of replication, towards novel functionality is explored. In the three polymerases created, modeling and large scale approaches are used to demonstrate the additional capability of each new enzyme. Next, two independent studies in the genomic adaptations needed for *E. coli* cells to adapt a 21st amino acid (selenocysteine and nitrotyrosine) are presented. Next generation sequencing is used to better understand the mechanisms of how cells accommodate the increased fitness burden placed by an orthogonal translation system. Lastly, community-wide changes in the oral microbiome are studied in the progression towards periodontitis, with implications towards potential therapeutic targets.

The capstone of this thesis leverages big data techniques to engineer novel proteins, the chief functional units within cells. Protein structural data is implemented into a

convolutional neural network to associate amino acids with neighboring chemical microenvironments at state-of-the-art accuracy. This algorithm enables identification of gain-of-function mutations, and subsequent experiments confirm substantive improvements in stability-associated phenotypes *in vivo* across three diverse proteins. This work is the first demonstration of using deep learning to empirically improve protein function and opens a new avenue for protein engineering.

Table of Contents

List of Figures xii

List of Tables xiv

Chapter 1: Introduction

Machine Learning and Proteins 1

1.1 Data Driven Biology 1

1.1.1 Introduction 1

1.1.2 Databases 2

1.1.3 Artificial intelligence 2

1.2 Computational methods for enzyme annotation 4

1.2.1 Generative approaches 4

1.2.2 Discriminative approaches 5

1.3 Machine learning and protein engineering 6

1.3.1 Linear models 6

1.3.2 Nonlinear models 7

1.3.3 Deep learning models 8

1.4 Deep learning representations for proteins 9

1.4.1 Sequence based inputs 9

1.4.2 Structure based inputs 11

1.5 Protein Folding 13

1.6 Generative networks 14

1.6.1 Variational auto-encoders 14

1.6.2 Autoregressive models 15

1.6.3 Generative adversarial networks 16

Chapter 2

A structure-based deep learning framework for protein engineering 17

2.1 Introduction 18

2.2 Results 19

2.2.1 Discrete modification improve wild-type prediction accuracy 19

2.2.2 Benchmarking neural network predictions 20

2.2.3 Neural networks identifies candidate sites to improve stability 20

2.2.4 Predicted substitutions can reduce screening load ... 21
2.2.5 3DCNN can unravel biological mechanisms ... 22
2.2.6 Comparison to other computational approaches ... 22
2.4 Discussion ... 23
2.5 Methods ... 23
2.5.1 Dataset and training ... 23
2.5.2 Confusion Matrix and Regression Bias ... 24
2.5.3 Rosetta/FoldX Calculations ... 25
2.5.4 Molecular biology ... 25
2.5.5 Protein Purification ... 25
2.5.6 BFP Fluorescence Assay and biophysical measurements ... 26
2.5.7 TEM-1 Assay ... 27
2.5.8 PMI Assays ... 27
2.5.9 NGS and variant calling ... 28
2.5.10 Statistical Methods and Data Presentation ... 29
2.6 Conclusions ... 29

Chapter 3

Directed Evolution of Polymerases ... 43
3.1 Introduction ... 45
3.2 Results ... 50
3.2.1 Variants from directed evolution can perform LAMP ... 50
3.2.2 Thermoresistance is enhanced in the evolved polymerase variant, v5.9 ... 51
3.2.3 v5.9 has novel hyperbranched RCA activity ... 52
3.2.4 RTX proofreads on RNA and DNA templates ... 53
3.2.5 RTX can streamline established RNA workflows ... 54
3.2.6 RTX can be further evolved towards xDNAs ... 54
3.3 Discussion ... 56
3.4 Methods ... 60
3.4.1 Polymerase purification ... 60
3.4.2 Real-time LAMP screening ... 61
3.4.3 Thermoresistance assay ... 62
3.4.4 Rolling circle amplification assays ... 62

3.4.5 Reverse transcriptase fidelity 63
3.4.6 RT-PCR Assay 64
3.4.7 RNA sequencing and analysis 64
3.4.8 Encoding of information into oligonucleotides 65
3.4.9 Synthesis of DNA and Omethyl DNA for Cryptogenetic Storage 65
3.4.10 Preparation of DNA for NGS Sequencing 66
3.4.11 Informatic recovery 66
3.5 Conclusion 67

Chapter 4

Bacterial adaptations towards the adoption of expanded genetic codes 88
4.1 Introduction 89
4.2 Results 90
4.2.1 Establishing selenocysteine dependence in E. coli 90
4.2.2 Genetic analysis of selenocysteine evolved strains 91
4.2.3 Experimental evolution of bacteria in the presence of 3nY 97
4.2.4 Genetic analysis of 3nY adaption 98
4.4 Discussion 102
4.5 Methods 105
4.5.1 Molecular biology 105
4.5.2 Bacterial passaging and growth assays 106
4.5.3 Whole genome sequencing and bioinformatic analysis. 107
4.5.4 qPCR and mutant allele detection 108
4.5.5 Statistical analysis and reproducibility 109
4.5.6 Evolutionary set up 109
4.5.7 Genome sequencing and assembly 110
4.5.8 Data availability 110
4.6 Conclusion 110
Figure 4.1 Evolution of selenocysteine-dependent E. coli strains improves fitness. 112

Chapter 5

Microbiota and Metatranscriptome Changes Accompanying the Onset of Gingivitis 125
5.1 Introduction 126
5.2 Results 128

5.2.1 Microbiota composition is correlated with clinical gingivitis index and patient 128
5.2.2 Shifts in relative genus abundance occur during health to gingivitis transition 129
5.2.3 Overall RNAseq transcript data correlates with 16S rRNA sequencing data 130
5.2.4 Virulence-related expression is elevated during the transition from health to gingivitis 132
5.2.5 Individual periodontal pathogens upregulate expression of both specific and general virulence-related genes during gingivitis relative to health 133
5.3 Discussion 135
5.4 Methods 139
5.4.1 Patient population and study design 139
5.4.2 Clinical assessment of gingivitis. 141
5.4.3 Cytokine analysis. 142
5.4.4 RNA isolation and preparation. 142
5.4.5 Bioinformatic analysis 143
5.5 Conclusions 145
References 160

List of Figures

Figure 2.1 Design and performance of a deep learning program capable of classifying wild-type amino acids with improved accuracy..... 30
Figure 2.2 Improvement in predictive accuracy of the model when training was normalized for amino acid abundance. 31
Figure 2.3 Classification accuracy with deep mutational scanning data..... 32
Figure 2.4 Precision recall curve for three computational methods. 32
Figure 2.5 Empirical validation of the model as a tool for protein engineering..... 33
Figure 2.6 Fluorescence for site-saturation libraries at disfavored residues in secBFP..... 34
Figure 2.7 Fluorescence for site-saturation libraries at random locations in secBFP 35
Figure 2.8 Fluorescence for site-saturation libraries at favored residues in secBFP..... 36
Figure 2.9 BFP-Bluebonnet exhibits improved folding compared to parental proteins..... 37
Figure 2.10 Mutant CaPMI variants complement deletion of the E. coli manA gene..... 38
Figure 2.11 Fluorescence assay of secBFP2.1 variants 39
Figure 2.12 Antibiotic resistance assay of TEM-1 β-lactamase variants..... 39
Figure 2.13 Fluorescence assay of the split-GFP-CaPMI fusions. 40
Figure 2.14 Masking of atoms reveals the mechanism of a global stabilizing mutation. 41
Figure 2.15 Venn diagram showing overlap between different computational tools..... 42
Figure 3.1 Schematic of HTI-CSR. 69
Figure 3.2 Isothermal screening of evolved variants..... 70
Figure 3.3 Nicked RCA reactions with variants isolated from HTI-CSR selections. 71
Figure 3.4 LAMP amplification with Bst LF and SD Pol..... 71
Figure 3.5 Thermoresistance kinetics of v5.9..... 72
Figure 3.6 Hyperbranched RCA with v5.9, Bst LF, and Klentaq..... 73
Figure 3.7 Evolution of a synthetic family of reverse transcriptases by RT-CSR. 74
Figure 3.8 EvaGreen qRT-PCR analysis using RTX 76
Figure 3.9 RTX polymerase proofreads during reverse transcription..... 77
Figure 3.10 The SSCS method for reverse transcription 78
Figure 3.11 RTX in an RNA-seq workflow..... 80
Figure 3.12 Evolution of a xDNA/polymerase pair creates a secure platform for DNA information..... 81
Figure 3.13 Encoding and decoding of information into oligonucleotides. 83

Figure 3.14 Information storage encoding schematic. 84
Figure 3.15 Distribution of NGS read sizes. 85
Figure 3.16 Decoding workflow of secured and unsecured oligos. 86
Figure 4.1 Evolution of selenocysteine-dependent E. coli strains improves fitness 112
Figure 4.2 Genes mutated during continuous evolution. 114
Figure 4.3 Growth curves of RTΔA cells containing point mutants in oxidative stress and selenite resistance. 117
Figure 4.4 Mutant *cysK* allele retention and plasmid copy number determination 118
Figure 4.5 Growth curves of C321.ΔA and *prfB* (release factor 2) single point mutants in RTΔA cells in rich and defined media. 119
Figure 4.6 Ceftazidime MICs. 120
Figure 4.7 3nY evolution set up. 121
Figure 4.8 Growth rates of parental and evolved strains 122
Figure 5.1 Study design and visualization of progression to periodontal disease. 147
Figure 5.2 Shannon index of within sample (alpha-diversity) of 16S rRNA sequencing reads from all samples. 149
Figure 5.3 PCoA analysis of subgingival plaque sample 16S rRNA sequencing reads. 150
Figure 5.4 PCoA clustering of 16S rRNA sequencing samples by changes in cytokine 151
Figure 5.5 PCoA clustering of 16S rRNA sequencing samples by MGI score and visit. 152
Figure 5.6 Composition of microbial communities from dental plaque samples assessed by 16S rRNA sequencing analysis 153
Figure 5.7 Simulated mapping analysis and mapping statistics. 154
Figure 5.8 Comparison of transcriptomic data between disease states, patients, and taxonomic abundance. 155
Figure 5.9 Genera and species comparison for 16S rRNA and RNA sequencing data 156
Figure 5.10 Genera abundance comparison between 16S rRNA and RNAseq data. 157
Figure 5.11 Virulence-related genes with significant differential expression 158

List of Tables

Table 3.1 Deep sequencing of RT-CSR libraries 75
Table 3.2 Fidelity for reverse transcription and replication. 79
Table 3.3 NGS sequencing of the Ome RT-CSR Round 18 pool 82
Table 3.4 Decoding trials 87
Table 4.1 Amber mutations in evolved lines. 113
Table 4.2 SNPs enriched in populations evolved under β-lactam stress. 115
Table 4.3 SNPs enriched in populations evolved under thermal stress. 116
Table 4.4 Mutations found in bla_{Addicted} and bla_{Control} in evolved lines. 123
Table 4.5 Genomic Mutations of Evolved Lines. 124
Table 5.1 Patient clinical plaque sample data. 148
Table 5.2 Upregulated virulence-related genes of representative periodontal pathogens from the five most abundant genera during gingivitis (MGI=2) 159

Chapter 1

Introduction
Machine Learning and Proteins

1.1 Data Driven Biology

1.1.1 Introduction

A new period of biology began in 2003 with the completion of a draft of the human genome. At the onset of this daunting task, it was believed that a curtain would be lifted from biology; the intricacies and complexities that have mystified molecular biology researchers would be unraveled. Like the Rosetta stone unlocking the secrets of early civilizations, so too would the genome provide the key to biology. But what happened was the exact opposite. Rather than shine light on the unknowns, the human genome draft opened up much more mysteries. The scientific community realized DNA is not the 'end all, be all' as originally thought, but a cog in a larger picture with transcriptional regulation, protein folding, and epigenetics among others all having an influence[1].

Where the human genome endeavor succeeded was in changing the biological narrative. Traditional biological approaches rely on interrogating a single gene or protein to gather a deep understanding of it behavior. However, in light of the human genome project, a bigger picture needed to be taken into account. A direct consequence of the draft human genome was the advent of next generation sequencing, which has fundamentally changed the way researchers attack their experimental hypothesis towards more data-driven approaches[2]. Sophisticated algorithms and tools can comb through very large amounts of data to derive knowledge from seemingly spurious information. This type of research forgoes an a priori hypothesis or assumption about what results may come out and instead relies on the data to form conclusions. At a first pass, this seems antithetical to the scientific method, where the cycle of formulating a hypothesis, designing an experiment to test, analyzing the results, and reformulating the hypothesis to explain the experiment is predominating. Instead, a bottom-up, inductive approach is used where ganged together facts are used to generalize meanings.

1.1.2 Databases

The explosion of data into biology can be highlighted by the sheer number of databases, both general and specialized, that have been created and, in turn, stimulated the breeding ground for machine learning tools. Presently, more than one thousand databases exist in biology addressing various different –omics approaches[3]. The biological information is synthesized in a standard format that eases readability and accessibility for machine learning researchers. The largest of these are classified into primary databases, which serve as computational archives with raw sequences of DNA or protein. Widely used examples include GenBank for genome sequences or Uniprot and the Protein Data Bank for protein sequences and structures, respectively. For tailored training tasks, secondary databases can be used which subdivide a primary database by an annotated feature. Of particular interest to the protein world are the SCOP and CATH databases which aim to classify subdomains of proteins based on structure and the PROSITE and Pfam databases where functional domains of various domains serve as the entries. Combining manually curated data is not without its challenges; differences and lab-to-lab variation can introduce unwarranted bias. As a solution, machine learning is employed to assess data quality and integrity, including Genbank and other widely used databases[4].

1.1.3 Artificial intelligence

Along with sophisticated data curation, the other enabler of machine learning in biology is the maturing of artificial intelligence algorithms. First, a discussion on the difference between machine learning and artificial intelligence is presented. An umbrella term, artificial intelligence, refers to the use of machines to enable capabilities in perception, logic, and learning. Machine learning is a subfield of AI uses algorithms to learn from data to make predictions, with the accuracy of the model increasing with more data used. Typically, feature extraction is performed by the user to extract the relevant information to be fed into the model. This type of learning encompasses regressions, support vector machines, decision trees, and principal component analysis. On the other hand, deep learning, a further specialization of artificial intelligence, use many layered neurons to build algorithms that are capable of learning important and relevant features on its own to boost performance on a

training task. Because features are not user generated and the nature of layered neurons, model interpretation into why it is performing how it is remains a challenge.

To give a more concrete example, imagine developing a model that could recognize a picture of a cat. In the machine learning approach, certain features of the cat would need to be first identified which could either be direct (notations for eyes, whiskers, fur, etc.) or metaphysical like a hypothetical cuteness index. The performance of the model would be tied to how well the extracted features generalize the idea of 'catness'. More data could enable tweaking of the model, but the general algorithm remains the same. This could potentially limit detecting different breeds of cats or discriminating cats from other animals. Contrary, a deep learning approach is agnostic to the any features of a cat. Instead, it leverages thousands to millions of images of cats to be able to learn the discerning features. As long as the input data is of sufficient quality and integrity and the training task is initialized correctly, then this type of approach is powerful to gather insight without knowledge of the underlying or important features. To connect towards protein design, an analogous approach has dominated the field. Modeling protein interactions with energy and statistical based potentials are manually extracted features that are used by most in the field. Chapter 2 takes the other tact and supposes proteins can be better engineered through deep learning.

Deep learning, surprisingly, is not a recent concept despite the recent hype and publicity. The idea of a neural network was first put forth by Frank Rosenblatt in the proposal of a perceptron[5]. Here, a neuron acts as a mathematical function and assigns a weight to the input. Multiple perceptrons are fed into a nonlinear activation function to capture complex phenomena. However, research was halted and entered an 'AI winter' as deeper networks lacked the necessary computing power. A revival occurred in 1986 with the idea of backpropagation[6]. Here, if a network misclassifies an input, then the weights are adjusted in the direction of the largest gradient such that a correct output is given if the input is shown again. Put more simply, the network is nudged in the direction to more correctly guess the desired output. This milestone enabled networks to be more computationally tractable, yet widespread adoption was still hindered by computational resources. In 2012, a network was reported on, AlexNet, that combined large computational power with these networks. In an image classification competition, this landmark model outperformed rival methods by 25%[7]. The key enabler was twofold; the use of graphical processing units (GPUs) enabled the

backpropagation which are fortuitously suited for performing the necessary, repetitive calculations. Second, the network made use of large training sets to fine-tune the 60 million parameters. Despite the black box nature and inability to process complex inputs not present in the training data, deep learning quickly spread. Biological applications will be discussed in subsequent sections.

1.2 Computational methods for enzyme annotation

1.2.1 Generative approaches

Novel technologies like NGS have exploded the number of observed sequences; the mining of genomes and metagenomics has become a viable strategy for searching for proteins with novel function. However, the majority of protein sequences are associated with an unknown function. For example, only 0.3% of discovered sequences have been manually curated with a function, indicating a need to identify the remaining 'dark' genes. Explicit experimental characterization of unknown enzymes is not feasible due to the scale and cost required. Exploring more tractable manners of annotation will expand the protein knowledge base to find proteins that could provide new avenues for biotechnology.

Traditionally, the practice of annotating proteins relied on alignment of sequences with known function. The first of these such methods was BLASTp[8], which uses pairwise alignments to find the degree of similarity to a homologous, (but crucially) annotated protein. Though in some sense crude, BLASTp is still predominantly used, however, the algorithm fails when insufficient homologs exist at roughly about < 30% similarity[9], thus limiting the utility towards not finding strikingly unique proteins but instead to compare against a known set. To illustrate this limitation, a study by Hess et al. utilized a metagenomics approach in cow rumen to identify biomass degrading enzymes and microbes adherent to plant fiber[10]. Of over 27,000 putatively identified enzymes, close to 43% of genes had < 50 % identity to any known protein in the NCBI database. Additionally, blast is computationally slow and scales linearly with search size. As the sequence world is growing exponentially, this method quickly becomes infeasible, underscoring the necessity to move away from homolog based searches.

The aforementioned and more traditional approaches to protein classification take a generative approach. A model is built for a single family or class of proteins, and a target sequence is evaluated for how well it fits that model. If the resulting score from the model is above a certain threshold, then that protein is labeled as such. The following sections describe a different tact using a discriminative approach. In this manner, proteins are labeled as positive or one for belonging to a class and negative or zero for not. The data representation then allows for a learning algorithm to infer the difference between classes as opposed to generative approaches which only necessitates a positive class.

1.2.2 Discriminative approaches

The first efforts into protein classification with a discriminative approach utilized support vector machines. This supervised method of learning draws a hyperplane through defined features to separate binary classes. In the protein classification problem, the first noted example was by Jaakkola et al., where a generative Hidden Markov model (HMM) was first used to establish a baseline model and extract similarity features from protein sequences[11]. A support vector machine algorithm is then applied to the set of feature vectors which was shown to outperform blast. However, because an HMM is used generating the features, the computational cost is limiting at roughly quadratic scale to the sequence length. Leslie et al. attempted to improve upon this work by eliminating the need for the generative step. Instead, features are provided in the form of sequence similarity with k-mers[12]. While the classification accuracy was slightly less with respect to less homologous protein families, the computational runtime was decreased to linear scale. Qiu et al. also explored using a support vector machine classifier but used structural feature kernels using the program MAMMOTH, outperforming other structural approaches[13]. Other machine learning methods have been ensembled together to further increase accuracy of classification[14].

Trending with the growth of deep learning, the next wave of work in protein functional prediction performed classification with neural networks, and most notably, without the need to explicitly identify features on which to train. A report by Cao et al. translated the k-mers of protein into an analogous language format, allowing the reformulation of the problem to be answered by a previously developed recurrent neural network and reported

comparable results to other methods[15]. Similar endeavors have successfully predicted gene ontology, enzyme classification, or remote homology detection[16–20].

In an interesting approach by Strodthoff et al., a similar recurrent neural network is utilized[21]. However, their model was first pre-trained on unlabeled protein sequences to learn generalities and makeups of proteins. Once in hand, the model was fine-tuned towards the classification task of protein function. The performance here outperformed state-of-the-art algorithms, highlighting that a more general approach may be the most fruitful. If inherent protein rules can be learned without a specific training task, then more data can be used to build this initial intuition in the model and provide a starting point for many diverse prediction tasks.

The above works suffer from an unavoidable bias that exists in biological research. Most annotated or curated proteins are derived from well-studied protein classes. To generate diversity into a training set, Wan et al. used a generative adversarial network called FFPred-GAN to learn the high-dimensionality features of protein space and apply that to augment their data set with high quality synthetic protein samples. Prediction accuracy for gene ontology was increased, highlighting a method where computer generated data can better inform training[22].

1.3 Machine learning and protein engineering

1.3.1 Linear models

Historically, the imparting of novel function or optimization into proteins is performed through directed evolution. In this experimental approach, multiple rounds of variant generation are selected for an intended function through screening. Despite recent success including the 2018 Nobel Prize in Chemistry, there is a need to make this process more efficient. Often, single mutations confer small, incremental improvements that in isolation are not enough for the threshold of success. Observed variants can be aggregated to identify improved variants, but selecting which mutations are truly beneficial and others that have hitchhiked (or may be deleterious) is cumbersome. Additionally, directed evolution inefficiently samples the total sequence space. As highlighted by Yang et al., for a 300 amino acid protein, there are 5,700 possible single amino acid substitutions and 32,381,700

possible variants with double mutations[23]. As this scale exceeds the capabilities of experimental approaches, more efforts in computation are being used to search sequence space in a more efficient manner.

The first foray to inform directed evolution experiments with machine learning used a simple linear model with a concept borrowed from drug development. Quantitative structure-activity relationship (QSAR) seeks to discover causal relationships between the structures of interacting molecules or physio-chemical properties with a measured response to optimize potential drug candidates. Fox et al. adapted this strategy to guide improved protein function by applying a linear transformation to the presence or absence of a mutation at a given position as the input features[24]. In validating the method, a 4,000 fold improvement was shown in a bacterial dehalogenase to produce ethyl (R)-4-cyano-3-hydroxybutyrate (HN). Extending on this work, Li et al. used a similar regression with features derived from sequence blocks in a chimeric library as opposed to single mutations[25]. Trying to improve the thermostability of P450, 184 chimeras were build and the melting temperature of each variant was measured. Their model was able to predict the contribution of each sequence fragment and generate novel chimeras with 108-fold increased thermal stability. Another report by Liao et al. evaluated 8 different regression models with different regularization methods to improve the activity of proteinase K. In designing and testing only 95 variants, a 20-fold increase in activity was observed[26].

1.3.2 Nonlinear models

These initial reports served to show that a simple model can streamline protein engineering experiments. Next, nonlinear models were explored soon after to capture added complexity not achievable by linear models with a particular focus on the use of Gaussian processes. This method is an alternative to regression models that creates probabilistic predictions by interpolating observations through Bayesian learning. In the protein context, this method aims to draw inference about unobserved mutants based on the distance from sampled mutants, thereby sampling greater sequence space without explicit measurements. Romero et al. first utilized this process to optimize proteins, using contacting residues as a kernel function to capture the distance between different sequences[27]. The model was able to correlate cytochrome P450 variants with different biophysical characteristics including

thermostability, enzyme activity, and binding affinity. In a similar manner, Bedbrook et al simultaneously optimized ChR for functionality, photocurrent amplitude, and light sensitivity[28]. Incorporating only 102 variants into the training set, this report showed how small amounts of data can be leveraged into a biologically useful new protein. Additional works show the utility of this method in improving fluorescence proteins by either shifting green fluorescence to yellow[29] or increasing green fluorescence[30].

1.3.3 Deep learning models

While these examples highlight the role of computational and statistical approaches to complement experimental evolution, it is limited in that the activity needs to be tied to a functional assay, affecting the throughput of variants that can be tested. In addition, Gaussian processes are unsuitable for large data ($>10^3$) as run time scales cubically with the number of training examples. Following the advent of next-generation sequencing, assays started to develop not towards a functional output but instead towards being tied to a sequencing output. The adaption of a NGS assay increases the number of variants that can be assayed and tested, thus creating data enrichment for deep learning methods. Deep learning tools are attractive as they require no foreknowledge of protein features, unlike in the previously described methods where the features needed to be explicitly stated. Khurana et al. reported the creation of a program, DeepSol which used convolution neural networks to predict protein solubility[31]. Using 58,689 known soluble proteins and 70,954 insoluble proteins to create a dataset, the framework extracted k-mer structure to perform binary classification of soluble/insoluble. Interestingly, raw sequence information was not enough to improve upon the state of the art methods. However, when explicit biophysical annotations were added into the model which included sequence features such as sequence length, molecular weight, alipathic index, average hydropathicity, charge, secondary structures, and solvent accessibility, the model outperformed all other classifiers. This work highlights how to build an accurate deep learning model in biology: local contextual filters learned by a convolutional neural network are complemented when aided by explicit structural and sequence annotations.

Elsewhere neural networks have been applied towards the thermostability prediction, benefitted by the use of the ProTherm database, a curation of experimentally derived

changes in temperature from mutations. This database contains data on 647 proteins with 22,713 variants tested[32]. When neural networks were in their relative infancy and yet to show broad application, Kwasigroch et al., created PopMusic to predict the effects of mutations on thermostability[33]. Here, a statistical potential function is generated that is governed by the fact that different functions show better ability at different parts of the protein (i.e. core vs surface). To improve upon their previous work, Dehouck et al. used neural networks to search and tune the parameters within their function[34]. When testing against the ProTherm database, a r squared value of 0.79 is obtained, beating previous attempts. Another developed method, NeEmo, used explicitly generated features to predict thermostability[35]. At a certain residue, evolutionary information from multiple sequence alignment and neighboring residues as well as biophysical environment are used for input features. Using a training set of 2300, a model was built with 5 hidden layer neurons. When outliers are removed from the test set (selected automatically as having the largest residuals when regressed against delta delta G), PopMusic shows the highest correlation in delta delta G. However, when the full dataset is used, NeEmo is the best performer. Though success has been shown, using a limited training set like the ProTherm databases biases results to only a handful proteins studied. Deep learning in biology excels when hundreds of thousands to millions of observed data are used with an associated labels. The next section will delve into representations of proteins as inputs to deep learning in a more global context.

1.4 Deep learning representations for proteins

1.4.1 Sequence based inputs

As the role of sequencing becomes more and more prevalent through biological research, the number of novel genes and sequences vastly increases. However, many of these new sequences are not associated with any particular property or context. To leverage the sheer number of unlabeled data, researchers have started to use deep learning frameworks that can be applied to where the property of sequences is not known. The prevailing hypothesis is that because evolution biases sequences towards a degree of fitness, the underlying signature in the sequence record can be extracted in a manner analogous to

natural language processing where a given word's semantics can be derived from the surrounding context.

Patrick Ng explored the utility of this concept by representing DNA bases as variable k-mers with the development of his model, dna2vec[36]. Common practice for this task is one-hot encoding, however, as the length of the sequence grows, the dimensionality exceeds exponentially. An alternative proposed with the method word2vec is word embeddings, where a simple 2-layer network is trained on a word and its surrounding context[37]. When applied to DNA, variable length k-mers (between 3 and 8 nucleotides) are used for training. Interestingly, this representation recapitulates the summing of two k-mer vectors is equal to nucleotide concatenation and correlates to Needleman-Wunsch sequence similarity.

DNA sequences abstract a level of information that is encoded into the protein sequence. To rectify this, Yang et al. used the k-mer embedding strategy to identify protein properties that are captured by their model[38]. A two stage training process is utilized. In the first, an unsupervised process is learning similar to the dna2vec approach described above where over 500,000 unlabeled protein sequences are used to train the latent embeddings. The second, supervised process aims to learn to predict the central k-mer given the surround k-mers using a Gaussian process regression. To highlight the performance of their model, the authors show improved ability to predict mutational effects for channelrhodopsin localization and cytochrome P450 thermostability over sequence representations.

The extension of this method towards a learning framework has been approached by three different reports. Alley et al. used a multiplicative RNN framework with 1900 nodes on 24 million sequences to construct a hidden state for each amino acid, taking into account all previous amino acids[39]. The model, termed UniRep, clustered sequences based on phylogenetics, separated sequences based on structural fold, and predicted mutations that improved stability. However, the model treated proteins as unidirectional but in the context of proteins compared to other domains, forwards vs reverse does not matter. Using models that took this into account were Rives et al. and Heinzinger et al. with bidirectional adaptions of the LSTM. The former uses a complex type of architecture with a bidirectional transformer and the ability to capture both long range interactions[40]. The latter is a natural extension of Alley et al. with the use of a bidirectional LSTM, learning the probability distribution over a

protein sequence[41]. Taken together, these reports unlabeled data can be leveraged to learn relevant biochemical properties and subsequently tailored towards more specialized tasks.

1.4.2 Structure based inputs

The structures of proteins are being solved at an exponential rate. The wealth of structural data lends itself to deep learning methods as the richness and quality of data increases. A particular area of concern is the so-called inverse protein folding, where the corresponding protein sequence is derived from a three dimensional fold. Implications of such a scheme could lead to smarter library design for protein engineering by reducing the search space of variants needed to be screened.

In attempting to achieve this task, Li et al. created a two-layer neural network that inputted protein structures and featurized local amino acids to 114 features with respect to backbone angles and energy-based parameters[42]. In determining the amino acid identity, the model achieved a classification accuracy of 30.3%. Interestingly, this report explored training against a second objective function based off PSSM to allow the model more flexibility in predicting similar types of amino acids, but did not appreciatively improve model performance. A second report by the same authors attempted to improve upon their initial results[43]. Additional features were included in the new model, accounting for density of atoms and alpha carbon angles and increasing the total extracted features by 54. With these additional model parameters, the accuracy of classifying amino acids was increased to roughly 34%. In investigating which features contributed to the increase in accuracy, it was discovered that the energy based amino acid profiles were the single biggest influence.

A strength of deep learning is the ability to learn the important, discriminative features without explicitly providing them in the model. Torng and Altman took this tack in their approach to this problem of identifying wild type amino acids given local structural context[44]. Convolutional neural networks, popular in image recognition, are adapted in this work to treat protein as static images and learn how environments associate with each of the twenty amino acids. In their approach, a 20 angstrom cube is centered on the amino acid of interest. Analogous to image representation where RGB values are stored in three channels, the atoms within the cube are voxelized into separate channels based on oxygen, carbon, nitrogen, and sulfur identity, generate a 4D tensor for each amino acid with size

4x20x20x20. The neural network leverages 3D convolutions to extract features in a space tailored towards protein representation. In training their model on 600,000 distinct environments with equal amino acid representation across 32,000 protein structures, a 42% accuracy is achieved. The deep learning results are compared to an approach similar to the previous example where protein characteristics are hard-coded as inputs. This program, named FEATURE and also developed by the Altman lab, extracts 80 physiochemical features to describe a singular amino acid. The authors report a 42.5% accuracy in their deep learning framework, a close to 20% increase in the same task when using manually curated features. To show utilty of their model, the effect of mutations are analyzed on mutant T4 lysozyme structures. For mutations which have empirically shown to be destabilizing and disrupt the resulting protein structure, the 3DCNN more likely (85% accuracy) predicts these sites as wild-type, connecting the output of their model to real world data. Further trying to explain the black box model, the gradient of the weights can be used to determine which atoms strongly favor the presence of the wild type amino acid. This work is the first convincing report to tie biochemical interactions to deep learning.

The majority of convolutional neural network applications relate to image processing, and thus using Cartesian coordinates makes translatable sense. As these sophicated networks are applied to more specialized fields, the representation of data may need to be changed in a way that normal operations no longer apply. Boosma and Frellsen posited that as atomic forces are predominantly distance based, than a radial parameteriztion may be more applicable[45]. Using the positional information, atomic mass, and the partial charge of each atom, the authors built a framework that utilized spherical convolutions on the input data. Two different types of convolution were explored, the first mapping coordinates into polar space and the second deconstructing a sphere into six neighboring cubes. In performance, both spherical representations had a higher accuracy than Cartesian representation, with the best spherical method outperforming by 6% and outperforming Torng and Altman by 8%, though it is difficult to draw a direct conclusion as the amino acid environment representations differ.

The previous two methods provided a consistent orientation to the input data, but the models and the learned filters likely will not transfer to amino acids with slight rotational movements. To accommodate this increased generalization, Weiler et al., proposed using

modified architectures that is equivariant to rigid body motions[46]. In this manner, learning is more efficient as symmetries can be captured that in previous examples would not be able to. In building the network as Boomsma save for the steerable 3D convolutional filters and replicating the training data, Weiler et al increased classification accuracy of amino acid by environemtn by 2%. While this represents a minimal step increase from the previous report, the framework is perhaps more suited to other protein prediction tasks, including interface and ligand binding.

1.5 Protein Folding

Another application of deep learning on protein sequences is the task to predict structural features. As the number of known protein sequences are orders of magnitude larger than the number of protein structures solved, this task is one of the unsolved challenges in molecular biology. A landmark study by Jinbo Xu using deep learning occurred during the 12th version of the CASP competition, a biannual competition to predict three dimensional folds from only a given sequence. The previous prevailing approach was to infer residue contacts from evolutionary couplings. Xu's point of departure was to instead predict distance prediction by way of a convolution neural network, changing from a binary classification (contact/no contact) to a discrete measurement[47]. In particular, distance prediction was divided into 25 bins spaced at 0.5 angstroms from 4 to 16 angstroms. Given an input sequence, multiple sequence alignments were used to derive sequence based and pairwise features. To further extract relevant features, a 1D network is first used to capture the sequence features of a residue or motif. The output is then combined with the related sequence information into a 2D network to learn pairwise contexts and eventual output a distance matrix for all atoms, secondary structure, and torsion angles with upper and lower bounds, which is then fed into a structure prediction algorithm.

Mohammed AlQureshi also took the approach of using deep learning to predict structures in the absence of co-evolution data[48]. Instead, a protein sequence with a PSSM derived from a multiple sequence alignment is inputted into a recurrent bidirectional neural network to output the torsion angles of a particular residue. The set of angles per residue is than aggregated across the entire protein by sequentially building the backbone of a protein

with each residue before taken into account. The last unit in the network uses global information to create a 3D structure by minimizing a RMSD function. AlQureshi found that his approach outperformed other models when predicting novel folds and was comparable to other template-based methods in predicting known folds.

Making waves in the next CASP competition was an entry by Google's DeepMind with their algorithm AlphaFold, where the distance to second place was greater than the past four competitions[49]. Building upon the work by Xu, the AlphaFold method predicts distance as the final component of the network at discretized intervals. However, it deviates by using the entire distribution rather than the mean and variance as a statistical potential borrowing energy terms from Rosetta that is minimized to predict protein fold. Powered by the expertise of the DeepMind team, their approach also leveraged complex networks hundreds of layers deep with computational tricks to enable longer training and vast hyperparameter space searching to bolster performance[50].

1.6 Generative networks

The discussion up to now has focused on classification tasks. More specifically, we have discussed the role that deep learning can identify anomalies associated with a given training task. An exciting and growing application of deep learning are generative models, which learn input representation and can output candidate proteins. Three different methods have been explored in relation to protein design—variational auto-encoders, autoregressive models, and generative adversarial networks.

1.6.1 Variational auto-encoders

Generative models generate random but new outputs that look similar to training data. In comparison with other types, variational auto-encoders are useful in exploring and varying the data already present in a specific desired direction. The hallmark of variational auto-encoders are two connected networks, an encoder and a decoder. The encoder takes an input and compresses it to a lower dimensional space with enough information to reconstruct an output. These latent encodings are trained to discriminate between protein features, thus providing the design space that can be transversed. Each latent variable has

an associated mean and standard deviation such that when generating new outputs through the inverse decoder network, the distribution can be kicked in a certain direction that makes sense in light of the inputs.

The first proof of concept of this type of approach was put forth by Sinai et al. By use of this method, they hypothesized that given an input sequence and its associated multiple sequence alignment, their model could encode relevant features and assign a likelihood that new sequences could be explained by their model and indeed show clustering of sequences based on phylogenetic distance and the ability to predict if a new sequence is functional[51]. Expanding upon this work, Riesselman, Ingraham, and Marks build a similar type of model with an increased input data set and the use of Bayesian variational approximation for weight optimization, and showed their model could predict mutational effects better than supervised methods[52]. Costello and Martin further expanded the scope by training their variational auto-encoder, called BioSeq VAE, on all protein sequences within the SwissProt database and show generation of sequences that can fold and are functional[53]. Further highlighting the power of variational auto-encoders, two approaches apply this method though specialized conditioning towards the design of T cell receptors, where a VAE model could learn the rules of VDJ recombination and generalize to unseen reportoires[54], and metalloproteins, where new metal binding sites are introduced into existing proteins[55]. Through these works, it is remarkable that a lower dimensional latent space can extract relevant protein features and be used to reconstruct new or altered proteins.

1.6.2 Autoregressive models

The previous models rely on homologous sequence alignment, which works well for well-defined protein families but not for those where not much information is known. To circumvent this constraint, another type of generative model for protein design is explored, autoregressive models. These models, commonly used to fill in gaps within images, attempt to predict the next input by taking into account every input that came before. In the context of proteins, this means that predicting residues with information from all preceding positions. Riesselman et al. employed this type of model to predict the effects of insertions and deletions in protein sequences, mutational types that are often ignored in protein design, and explained mutational effects within a therapeutically relevant nanobody[56].

1.6.3 Generative adversarial networks

One limitation in these methods is the enforced directionality of protein sequences. Generative adversarial networks are the next algorithms applied to generate *in silico* proteins that circumvents this constraint. In this type of deep learning architecture initially proposed by Goodfellow, two separate networks are created, a generator and a discriminator[57]. The generator network creates synthetic, realistic data by learning underlying characteristics from training data. The generated data is then fed into a discriminator network which aims to discriminate between real data and the data produced by the generator network. To iteratively improve performance, the loss of generator network is tied with the discriminator network such that increasingly more realistic data is generated. A perfect generative adversarial network would have an end result of the discriminator guessing randomly, thus the inability to distinguish real from generated data. In practice, generative adversarial networks are tricky to implement as they are sensitive to hyperparameters and requiring a delicate balance to train correctly. In Gupta and Zou, a generative adversarial network is created to generate small peptides with antimicrobial properties[58]. In their architecture, the top scoring peptides are kept while low scoring peptides are dropped from the training set, and the generator gets tasked to mimic these top results. While novel peptides were generated in this framework, the retention of antimicrobial activity is not shown. Additionally, the authors trained on DNA sequences rather than amino acid composition, thus abstracting a level of crucial information. In a second instance, Karimi et al., used a modified generative adversarial network to create novel protein folds[59]. They find that the use of a generative adversarial network outperformed variational autoencoders in both generating increased sequence diversity in protein folds and have better fold accuracy. Though in relative infancy, generative adversarial networks represent an intriguing new method to learn and generate novel protein features.

Chapter 2

A structure-based deep learning framework for protein engineering

The first chapter brings the forefront of artificial intelligence to synthetic biology. After observing the necessity of stabilizing mutations to promote new enzyme function as in Chapters 3, 4 and 5, we pondered if computation could be used to increase the overall stability of a protein before being subjected to selection. Most functional mutations tend to destabilize a protein; therefore, if the thermodynamic cost for the new mutation impedes the overall folding energetics, then the mutation will not fix and not be observed in a population. As most natural proteins exist to be only marginally stable, the threshold between a folded and unfolded protein can be razor thin. Hence, we used a data driven approach to find candidate sites in a protein that can increase overall stability with a downstream goal of increasing the chance of success for a selection.

To address this, we use convolutional neural networks, which have recently become the preferred AI solution to a number of image recognition challenges, but are underutilized in biological applications. Proteins are exponentially being crystallized to solve their three-dimensional structure into publicly available repositories. By making the analogy of image pixels to atomic coordinates, we utilize a 3D convolutional neural network to engineer proteins with improved stability associated phenotypes. Neural network training is performed on 1.6 million local environments surrounding a central amino acid to learn a structural consensus for every residue. When then applied to a protein of interest, the model can identify positions that deviate from this consensus and serve as candidates for mutagenesis to improve protein function. The strength of this work is underpinned by the large training set which allows the model to be generalizable and associate local

This chapter is adapted from a draft manuscript Shroff R, Cole AW, Morrow BR, Diaz DJ, Donnell I, Gollihar J, Ellington AD, Thyer R. (2019).I shared first authorship with AWC and conceived the project, developed the code, designed and performed experiments, and wrote the manuscript.

environments with certain amino acids. These results forecast new biological tools at intersection of AI and molecular biology.

2.1 Introduction

Protein engineering is a transformative approach in biotechnology and biomedicine commonly used to alter natural proteins to tolerate non-native environments[60], modify substrate specificity[61], and improve catalytic activity[62]. Underpinning these properties is a protein's ability to fold and adopt a stable active configuration. Currently, stability is engineered from sequence information by identifying mutations accrued through evolutionary drift and reverting those mutations to homologous consensus residues[63], or alternatively, from structural information by simulating the dispersed chemical interactions throughout a protein to calculate the aggregate energetic effects of specific substitutions[64]. These structural methods vary in run time but are generally computationally expensive. Hydrophobic core repacking for a single substitution takes minutes, MD simulations for a single substitution take tens of minutes, and QM/MM methods take hours or days[65–67]. Alternatively, deep learning approaches taking seconds to minutes have been reported, however most models either predict empirically measured stability effects in biased datasets containing only thousands of annotated observations[68] or require model training on the target protein[39,69]. Nonetheless, machine learning models present fast and accurate approximations to protein structural feature prediction and assessment without requiring structural simulations.

Recently, three-dimensional convolutional neural networks (3D-CNNs) have been used to predict ligand pocket affinities[70], categorize proteins[71], and classify amino acids provided the surrounding local atomic environment[44]. In the latter application, a 3D-CNN was trained using high-resolution crystallographic data to map local protein microenvironments to their central amino acid and then asked to analyze every microenvironment-residue pair in a protein of interest. When presented with the local environment of empirically assayed positions, this model is able to identify wild-type residues at positions which disfavor substitutions and could assign a stable wild-type residue when presented with known destabilizing mutations[44]. With this framework, we sought to develop a novel approach for

engineering protein stability; we hypothesize that residues where the native amino acid is strongly disfavored are destabilizing such that protein folding and stability can be improved upon mutagenesis.

2.2 Results

2.2.1 Discrete modifications improve wild-type prediction accuracy

As a starting point, we rebuilt the neural network architecture published by Torng and Altman with minor modifications (**Figure 2.1a**) and successfully replicated the reported classification accuracy of 41.2% using the original training and testing sets (32,760 and 1601 structures, respectively)[44]. To improve the model's performance of associating amino acids with their environment, we made several discrete changes towards more explicit biophysical annotations (**Figure 2.1b**). First, we added a new atomic channel containing the coordinates of hydrogen atoms, modestly increasing accuracy to 43.4%. We next added additional biophysical channels to accommodate the partial charge and solvent accessibility. Addition of these channels increased wild-type prediction to 52.4%.

The sampling methodology for both protein structures and amino acid residues used to construct the original dataset was observed to introduce bias which resulted in non-optimal training data. The dataset contained multiple structures of closely related proteins which biased training towards overrepresented protein structures, where the 32,760 PDB IDs map to only 11,418 UniProtKB IDs. To improve dataset composition and uniformity, we gathered all PDB structures with less than 2.5 angstrom resolution and at most 50% sequence similarity, resulting in a training set consisting of 19,436 structures and 300 additional structures for out-of-sample testing. This improved dataset increased wild-type classification accuracy to 61%. We next wished to improve data consistency; deposited crystallographic structures are refined by algorithms of their time which are not necessarily the current state of the art. By drawing from structures in the PDB-REDO database, where existing protein structures are refined in a uniform manner, we increased accuracy to 63%.

The original dataset geometrically sampled amino acids which heavily biased training environments for surface residues. We removed this bias by sampling amino acids randomly throughout each protein which increased wild-type classification accuracy to 66%.

Additionally, the original dataset represented each amino acid at equal frequencies thereby biasing the expectation of the neural network towards rare amino acids (**Figure 2.2**). Sampling was altered such that amino acid frequencies mirrored their relative abundance in the PDB. The new training set, consisting of 1.6 million amino acid environments, improved classification accuracy to nearly 70%; this represents a new state-of-the-art accuracy for amino acid assignment versus previously published deep learning programs[42–44,72].

2.2.2 Benchmarking neural network predictions

To understand where errors from the neural network arose, we constructed a confusion matrix (**Figure 2.1c**). We noticed that similar amino acids are commonly misclassified, which suggests that the neural network recapitulates known biochemistry. Furthermore, proline and glycine, which have unique structures, are classified with above 96% accuracy while glutamine is classified at only 33% accuracy. We hypothesized that modifications increasing classification accuracy would disproportionately increase accuracy for amino acids that are well-matched to their environment. We tested this using previously published deep mutational scanning (DMS) data for the proteins TEM-1 β-lactamase, immunoglobulin binding domain of protein G (gb1), Aminoglycoside-3′-Phosphotransferase-Ia, ubiquitin, and Hsp90[73]. In this dataset, the effects of all possible single substitutions were quantified with a ceiling for activity set at wild-type function, i.e. no beneficial mutations were observable. We identified 292 positions where any substitution incurred a measurable fitness cost and benchmarked classification accuracy on this subset. The final version achieves a recall of 87.0%, signifying a 25.4% increase over the starting model (**Figure 2.3** and **Figure 2.4**).

2.2.3 Neural networks identifies candidate sites to improve stability

After establishing our ability to accurately classify wild-type amino acids, we next investigated where the wild-type expectation and our model most disagreed. Presuming the chemical and structural associations between amino acids and their microenvironments have been learned, then wild-type amino acids that are assigned a low probability to occur in their native environment could be substituted to increase conformity with the proteome and improve stability. We tested this hypothesis by building saturating libraries in

secBFP2.1, an engineered blue fluorescent protein[74], at residues assigned either the lowest (disfavored) or highest (favored) wild-type probabilities by our model. We also mutagenized ten residues selected at random to establish a control. Six of nine disfavored residues, one of ten random residues, and zero of ten favored residues could be substituted to improve fluorescence of secBFP2.1 ($p = 0.01$ by a Fisher's exact test for disfavored versus random subsets; **Figure 2.5a** and **Figures 2.6 – 2.8**). We amalgamated the beneficial substitutions into a single variant, designated BFP-Bluebonnet (BB), which improved florescence in *E. coli* by more than six-fold (**Figure 2.5b-c**). Furthermore, purified BFP-Bluebonnet exhibited improved thermal tolerance and chemical stability in guanidinium as compared to both secBFP2.1 and mTagBFP2 (**Figure 2.9**). To verify that our model was generalizable, we built site-saturation libraries at the ten most disfavored residues in each of two structurally and functionally unrelated enzymes, TEM-1 β-lactamase and *Candida albicans* phosphomannose isomerase (*Ca*PMI). Seven of the ten residues in TEM-1 β-lactamase and six of the ten residues in *Ca*PMI could be substituted to improve phenotypes associated with folding and stability (**Figure 2.5d-e**). Aggregating these stabilizing mutations improved the properties of *Ca*PMI by five-fold without abolishing catalytic activity (**Figure 2.10**).

2.2.4 Predicted substitutions can reduce screening load

While site-saturation mutagenesis at candidate residues is a good option for identifying beneficial mutations, it relies on the ability to screen protein variants in at least moderate throughput. To simulate a situation in which screening at such a scale is not possible, we examined the ability of the model to directly identify beneficial substitutions at residues where it did not assign the highest probability to the wild-type amino acid. We built all unique single point mutations in the three proteins (secBFP2.1, TEM-1 β-lactamase and *Ca*PMI) using the top ten substitutions generated by three different but largely overlapping interpretations of the model: the amino acid assigned the highest probability at the residues with lowest assigned probability of the wild-type amino acid (the residues selected for site-saturation mutagenesis), the amino acid assigned the highest probability where it differed from wild-type (regardless of the wild-type probability), and mutation to the amino acid assigned the highest probability differing from wild-type resulting in the greatest log-fold change over the wild-type probability (**Figures 2.11 – 2.13**). Using this approach, several

individual stabilizing mutations were identified for each protein and the effects were largely additive when combined. Although no single interpretation of the output data was clearly superior, this methodology resulted in at most 22 unique variants, which is a manageable number to synthesize and screen for all but the most challenging proteins.

2.2.5 3DCNN can unravel biological mechanisms

Although we focused on protein engineering applications for our model, it also has considerable potential as a tool to unravel fundamental biology. In particular, we sought to explain the model's ability to flag the mutation M182T in TEM-1 β-lactamase, a global suppressor mutation that has been identified in many clinical isolates. Despite its identification decades ago, the mechanistic explanation for stabilization remains under debate. One model proposes that the threonine hydroxyl forms an N-cap H bond with Ala 185 as determined through crystallographic analysis[75], while a competing explanation determined through molecular modeling suggests a stabilizing hydrogen bond with Glu 63 and/or Glu 64[76–78]. To find the contributing atoms that most favor a mutation to threonine, we systematically deleted every atom in the Met 182 microenvironment and used the used the model to analyze where the probabilities changed the most. Our method flagged two atoms, the backbone oxygen of Glu 63 and the amide hydrogen on Ala 185, in which the removal of either atom decreased the probability of observing a threonine by over 200 fold (**Figure 2.14**). Thus, a neural network framework can be used to suggest stabilization mechanisms in addition to identifying candidate residues for mutagenesis.

2.2.6 Comparison to other computational approaches

Two well-documented, alternative computational approaches to guide protein stabilization are Rosetta pmut_scan and FoldX PositionScan, both of which rely on energetics simulations. If our model learned inferences accessible by energetics calculations in either of these programs, we would expect significant overlap between the disfavored residues it identified and non-optimal positions predicted by either of these programs is expected. Only three of thirty positions identified by the model were also identified by either Rosetta or FoldX, which also largely identified separate residues. Furthermore, in TEM-1 β-lactamase, each of these methods uniquely identified stabilizing mutations reported elsewhere in the

literature (**Figure 2.15**). Therefore, our model can identify novel stabilizing loci not captured by other commonly used programs.

2.4 Discussion

Here we report a modified 3D CNN architecture with state-of-the-art classification accuracy for assigning wild-type residues throughout proteins. Where native amino acids deviate from their structural and chemical consensus, we demonstrate that these positions with low wild-type favorability are excellent targets for site-saturation mutagenesis and yield stabilizing mutants at frequencies that exceed random selection. Combining the stabilizing mutations identified in three model proteins improved variant phenotypes several fold relative to their ancestor. Furthermore, this model is synergistic with existing protein design tools by identifying sets of mutations that do not overlap with those derived from energetics simulations. This work is the one of the first demonstrations of using deep learning to empirically derive novel protein function and opens a new avenue for protein engineering.

2.5 Methods

2.5.1 Dataset and training

To reduce any bias resulting from the differential abundance of protein families in the PDB, we sought to build a dataset of protein structures with balanced phylogeny. To achieve this, we took all structures in the PDB database and clustered to 50% similarity to avoid oversampling towards certain protein classes. We further reduced the variability in the dataset by cross-referencing the structures to the PDB-redo database[79], which uses a consistent algorithm to refine, rebuild, and validate structures from raw crystallographic data. Within each clustered set of sequences, we identified the structure with the lowest resolution. If no structure existed below a resolution of 2.5 angstroms the entire cluster was discarded. This process yielded 19436 structures, of which 300 were randomly set aside for out of sample testing and the remainder used to generate the training set.

In addition to atomic annotations, our model adds additional channels for the partial charges and solvent accessibility associated with each atom. While all structure files label

oxygen, carbon, nitrogen, and sulfur, hydrogens may be missing depending on the resolution of the structure. Using the program pdb2pqr (v2.2.1)[80], hydrogens were placed into the structure and optimized while partial charges were assigned with the CHARMM force field. Solvent accessibility was calculated with the program FreeSASA (v2.0.2)[81]. To avoid oversampling residues from larger proteins, we limited the number of sampled environments from an individual protein to either half of the length of the protein or 100 amino acids, whichever number was less. Atomic environments consisting of a 20 angstrom cube centered around a single residue were generated as described in Torng and Altman[44].

The convolutional neural network was built using theano (v1.0.3) and consists of six layers, all with ReLu (rectified linear unit) activations. The first two convolutions were performed with a filter size of 3x3x3 with no padding and increased the depth to 200 channels. We then performed a max pooling step, followed by two additional convolutions with a filter size of 2x2x2 and increasing the depth to 400. Max pooling was used again before flattening and feeding into two successive fully connected layers with dropout rates of 0.5 and 0.2, respectively. Softmax activation was applied to the logits to obtain probability scores for each of the 20 amino acids.

Neural network training was performed on TACC's Maverick cluster with a NVIDIA Tesla K40 GPU. 1.6 million amino acid environments were generated with the abundance of individual amino acids mirroring the natural frequency observed in the PDB. As the dataset was too large to load entirely into memory, we split the data into 20,000 samples and randomly shuffled the order after loading. Batch sizes of 20 samples were used and the loss was calculated through RMSprop. Training was performed with an adaptive learning rate and lowered by 10% if validation accuracy did not decrease within 8000 training iterations. Four epochs were run, at which point overfitting was observed. Test and validation accuracy were measured in 6000 amino acid environments with equal representation of each residue.

2.5.2 Confusion Matrix and Regression Bias

To calculate the frequency at which wild-type residues were correctly predicted, 20,000 amino acid environments were generated from out of sample PDBs (i.e. structures not seen during training) with an amino acid distribution mirroring natural frequencies. Regressions highlighting amino acid bias were created by plotting the sum of the predicted probability

values against the frequency in the test set. The confusion matrix was generated by plotting the single amino acid assigned the highest probability at each microenvironment sampled compared to the wild-type amino acid.

2.5.3 Rosetta/FoldX Calculations

The pmut_scan program within the Rosetta software suite (v3.9) was used to calculate the computational effect of mutations with a large ΔΔG cutoff value to output both stabilizing and destabilizing mutations. To perform the analogous operation in FoldX (ver. 4), the PositionScan module was used. In either program, the least favorable sites were found by summing values less than zero (the sign change of a stabilizing mutation) and identifying the ten sites with the most negative value.

Recall and precision for each computational method was assessed through deep mutational scanning data sets from the corresponding structures: TEM-1 β-lactamase, PDB:1BTL; protein G, PBD:2QMT; aminoglycoside-3'-phosphotransferase-IIa, PDB:1ND4; ubiqutin, PDB:4XOF, and Hsp90, PDB:2BRC. Normalized fitness values were derived from Gray *et al.* (2018)[73] with a threshold of 1.02 to determine if a variant greater than wild-type exists. Within this subset, a positive result was defined if no other variant empirically exhibited better fitness than wild-type and, for our model, the wild-type amino acid was assigned the largest probability, or, for Rosetta and FoldX calculations, the minimum ΔΔG value (i.e. the most stabilizing value) was greater than zero.

2.5.4 Molecular biology

Experiments described in this manuscript were performed using standard molecular biology techniques. Unless otherwise indicated, all plasmids, single point mutations in reporter genes and site-saturation mutagenesis libraries were constructed using Gibson assembly. For site-saturation mutagenesis libraries, 2 μL of the reaction mixture was transformed into 50 μL of chemically competent *E. coli* cells. Transformations were required to exceed 10-fold library coverage (> 320 single colonies).

2.5.5 Protein Purification

To purify secBFP2.1 mutants, a 6xHis tag was appended to the C-terminus via a Gly-Ser-Gly linker. BL21 DE3 cells were cultured in Superior Broth to mid-log phase (~ OD_{600} 0.6) and induced with 1 mM IPTG for 16 hours at 18 ºC. Following induction, cells were harvested by centrifugation and lysed by sonication in 50 mM sodium phosphate, 300mM NaCl, 20mM Imidazole pH 7.4 buffer containing protease inhibitor (Pierce Protease Inhibitor) and Benzonase Nuclease (EMD Millipore). Cell lysate was clarified by centrifugation (40000 x *g*) and BFP variants purified using HisPur™ Ni-NTA Resin. Purified protein was dialyzed into 50 mM sodium phosphate pH 7.4 buffer and analyzed by SDS PAGE to assess purity.

2.5.6 BFP Fluorescence Assay and biophysical measurements

SecBFP2.1 was cloned into a kanamycin resistant derivative of plasmid pQE flanked by a T7 promoter and terminator. Site-saturation libraries were transformed into *E. coli* strain BL21 DE3 and a series of 10-fold dilutions (spanning two orders of magnitude) were plated on solid media to ensure sufficient discrete single colonies. 96-well deep-well plates were inoculated with 92 individual library transformants and four wild-type controls. Two plates were assayed for each library. Cells were cultured ON at 37 ºC in plate shakers at 850 rpm. 20 µL of the ON cultures were diluted into 880 µL LB and incubated for two hours. Cells were induced by the addition of 100 µL media containing 0.5 mM IPTG, resulting in a final concentration of 50 µM. After a four hour induction, cells were harvested by centrifugation and resuspended in 1 mL PBS. Fluorescence was measured on a Tecan M200 Pro using 400 nm for excitation and 460 nm for emission. A maximum of 12 individuals at each library site exhibiting fluorescence / OD_{600} values greater than wild-type were sequenced. Candidate mutations were re-cloned into the pQE plasmid and rephenotyped. Rephenotyping was performed in biological and technical triplicate.

Purified blue fluorescent proteins were diluted to 0.01 $mg.mL^{-1}$ in PBS pH 7.4 and 100 uL aliquots were heat treated for 10 minutes in PCR strips on a thermal gradient using a thermal cycler. Fluorescence of thermally challenged variants and controls incubated at room temperature was assayed using excitation and emission wavelengths of 402 nm and 457 nm respectively. Fluorescence readings were normalized to the mean of solutions incubated at room temperature e.g. a measurement of 0.8 indicates that a heat treated protein retained 80% of its untreated fluorescence.

Purified blue fluorescent proteins were diluted to 0.01 mg.mL^{-1} in 6 M guanidinium hydrochloride. 100 uL aliquots in technical triplicate were added to wells of a 96-well clear-bottom black-walled plate and incubated at 25 ºC for 23 hours. These purified fluorescent proteins were assayed at 30 minute intervals using excitation and emission wavelengths of 402 nm and 457 nm respectively. Plates were agitated preceding each measurement. Fluorescence values measured at time zero were used to normalize fluorescence through the remainder of the assay e.g. a measurement of 0.8 indicates that the protein retained 80% of its initial fluorescence.

2.5.7 TEM-1 Assay

The *bla*$_{\text{TEM-1}}$ gene encoding TEM-1 β-lactamase, including the native promoter, was amplified from pETDuet-1 and cloned into pCDFDuet-1 immediately upstream of the second T7 terminator, replacing both T7 promoters and both polylinkers. The L250Q mutation was introduced into TEM-1 to destabilize the protein and enable easy identification of compensatory stabilizing mutations[82]. Site-saturation libraries were transformed into *E. coli* strain DH10B and recovered ON in liquid medium supplemented with spectinomycin. ON cultures were diluted and plated on a range of different carbenicillin concentrations (0, 50, 125, 250 and 500 μg.mL^{-1}). For each library, 12 single colonies from the plate containing the highest concentration of carbenicillin were isolated and the *bla*$_{\text{TEM-1}}$ gene sequenced. Beta-lactamase variants identified by library screening were recloned into pCDFDuet-1 and rephenotyped. Rephenotyping was performed by diluting overnight cultures, in biological triplicate, 100-fold and spotting 5 μL onto solid media containing a gradient of carbenicillin concentrations.

2.5.8 PMI Assays

Improved variants of *Ca*PMI were identified using the split GFP reporter system described by Cabantous *et al*. (2008) with minor modifications[83]. Briefly, a fusion protein consisting of residues 173-238 of folding reporter GFP, a (GGGS)2 linker, residues 2-440 of *Ca*PMI, a (GGGS)2 linker and residues 2-172 of superfolder GFP was assembled in a derivative of pACYCDuet-1 (**SI Table 1**). Site-saturation libraries were transformed into *E. coli* strain BL21 DE3 and a series of dilutions plated on solid media supplemented with 0.25

mM IPTG. Following ON incubation at 37 ºC, plates were further incubated at 4 ºC for eight hours at which point highly fluorescent colonies were manually selected. PMI variants were subcloned and the fusion protein ORF fully sequenced prior to rephenotyping to ensure that increased fluorescence was not the result of mutations in the GFP fragments, linker regions, or plasmid backbone. Transformants were screened as described for BFP in 96-well deep-well plates in biological and technical triplicate. Fluorescence was measured on a Tecan M200 Pro using 475 nm for excitation and 535 nm for emission.

The *manA* gene encoding phosphomannose isomerase was disrupted in *E. coli* strain BL21 DE3 using lambda red recombineering to introduce a kanamycin resistance marker. Successful deletions were confirmed by colony PCR of Kan^R colonies using primers which flanked the *manA* locus. The wild-type *C. albicans manA* gene or variants containing combinations of stabilizing point mutations were cloned into a derivative of pACYCDuet-1 using Gibson assembly. BL21 DE3 Δ*manA*::kan cells were transformed with PMI expression plasmids and plated on LB agar with appropriate antibiotics. Single transformants, in biological triplicate, were transferred to liquid M9 minimal medium with 0.4% glucose and cultured ON. Cells were washed in a 1:1 volume of M9 medium without any carbon source and 2 µL streaked on M9 minimal medium plates supplemented with 0.4% mannose and 0.25 mM IPTG. Wild-type BL21 DE3 cells and BL21 DE3 Δ*manA*::kan cells containing an empty expression plasmid were used as positive and negative controls respectively. Plates were incubated at 37 ºC for 24 hours.

2.5.9 NGS and variant calling

Purified plasmids encoding BFP variants were quantified using the QuantIT dsDNA Assay Kit (Thermo Scientific) and 50 ng of each plasmid was pooled by well position, resulting in 192 samples each with a different variant for each tested position. Samples were prepped for sequencing by amplifying two discrete ~350 bp regions of the BFP gene with primers containing Illumina adapters and dual indexes. Sequencing was performed using Illumina MiSeq with paired end 2x300 bp reads.

Following sequencing, paired end reads were joined together using fastq-join (v1.3.1). Raw sequences were aligned to the original gene sequence with bwa (v0.7.17). Read counts across a single position were normalized to the observed fraction of each codon. Variants

that contained at least 100 read counts and exceeding 1% of wild-type counts were labeled as such, while variants that failed were left unlabeled. Outliers were identified through the OPTICS algorithm in the scikit-learn package (v0.21.3). Each sample was normalized by dividing the fluorescence by the OD_{600} and as well as to the average wild-type value per plate. Significance was determined using a Fisher's exact test where success of a position was defined as the mean fluorescence of the most fluorescent variant being at least three standard deviations higher than the wild-type mean.

2.5.10 Statistical Methods and Data Presentation

All data in the manuscript are displayed as mean ± s.e.m. unless specifically indicated. Bar graphs, regressions, confusion matrix, NGS variant graphs were plotted in R 3.4.1 using the package ggplot2 (v2.2.1).

2.6 Conclusions

As the wealth of biological resources grows, advanced computational tools will be increasingly common to make sense of data. We show here that one such resource, the Protein Data Bank, is amenable a deep learning architecture. While the results presented here are the first to use deep learning without *a priori* mutational knowledge to improve a protein, it is still unclear why the sites are being predicted. This is not unexpected as neural networks act as a black box between input and output, but there is potential to learn fundamental biology if the in-between layers could be parsed. Ongoing work is delving into this very notion, and could provide insight to better models and eventually a generative algorithm.

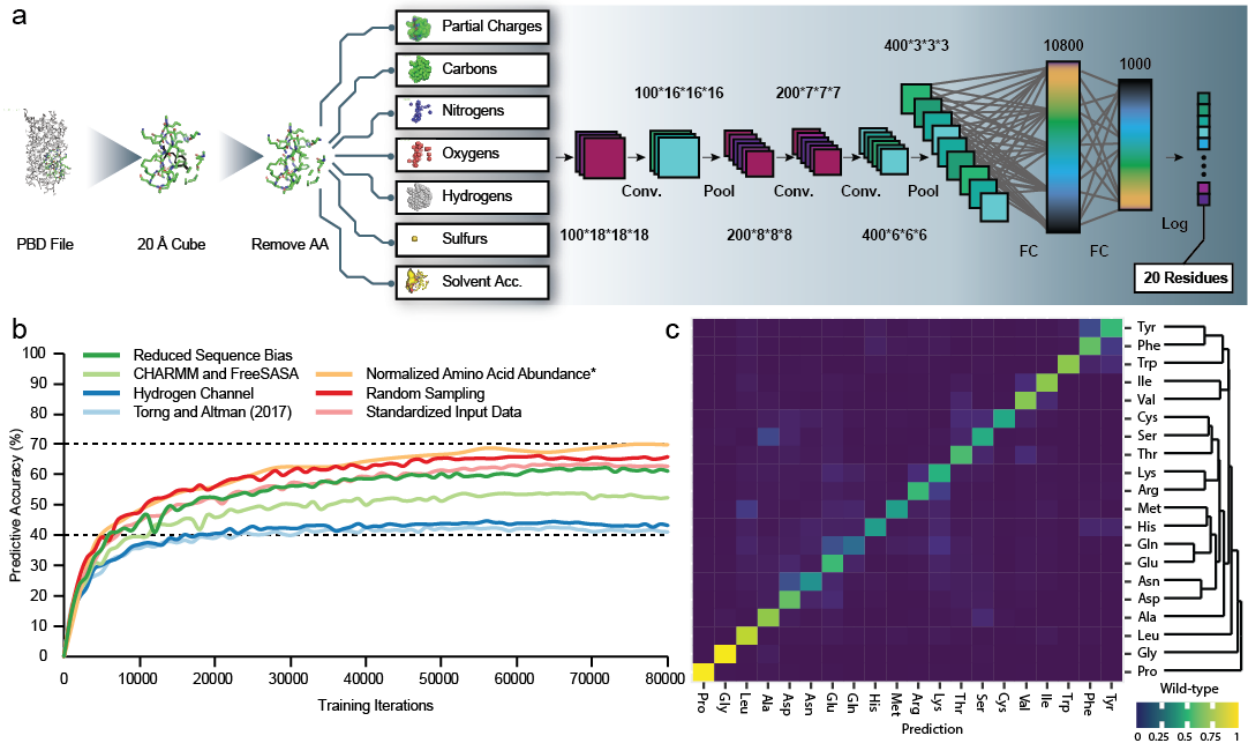


Figure 2.1 Design and performance of a deep learning program capable of classifying wild-type amino acids with improved accuracy.

a) Schematic of the model depicting the data pipeline and neural network architecture. b) Discrete changes made to the neural net framework described by Torng and Altman[14] and their effect on classification accuracy. *Normalizing the amino acid abundance of the training data increased the size of the dataset by roughly 4-fold. While the number of epochs decreased, the number of training iterations needed for convergence remained similar to the other versions. c) Confusion matrix showing bias of wild-type amino acid classification. Structurally unique amino acids Gly and Pro are assigned as wild-type with very high probability.

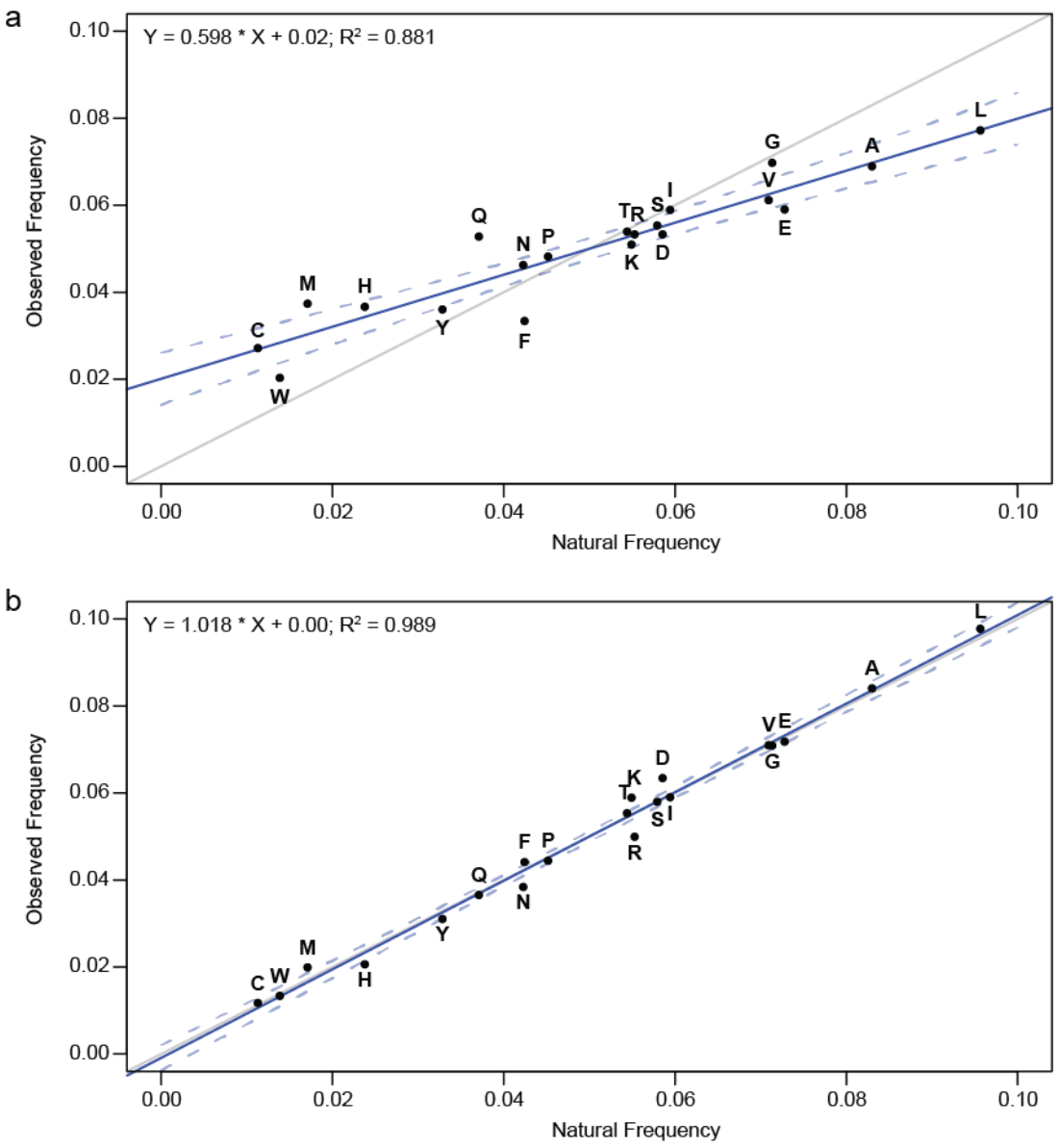


Figure 2.2 Improvement in predictive accuracy of the model when training was normalized for amino acid abundance.

a) Classification accuracy of the model trained for 84,000 iterations without normalization. The p-value for a two-sided t-test against null hypothesis: 'the slope is equal to 1' is less than 10e-6. b) Classification accuracy of the model following training for 240,000 iterations correcting for amino acid abundance. The grey line depicts a line with slope 1 and the blue line is the regression for the observed amino frequencies compared to the natural abundance. Dotted lines delineate 95% confidence intervals for the regression. The p-value for a two-sided t-test against null hypothesis: 'the slope is equal to 1' is 0.48.

Protein	PDB	n	4-Ch[5]	This work	Rosetta	FoldX
TEM-1 β-lactamase	1BTL	110	0.627	0.936	0.773	0.440
Protein G	2QMT	17	0.529	0.941	0.882	0.471
Aminoglycoside-3'-phosphotransferase-IIa	1ND4	77	0.727	0.870	0.610	0.493
Ubiquitin	4XOF	30	0.567	0.767	0.467	0.172
Hsp90	2BRC	58	0.500	0.776	0.569	0.379
Combined	-	292	0.616	0.870	0.664	0.426

Figure 2.3 Classification accuracy with deep mutational scanning data.

Accuracy of different computation tools for protein engineering using a dataset of true positive wild-type residues. For the 4-channel model and the model presented in this work, classification was considered correct if the wild-type amino acid was assigned the highest probability. For Rosetta and FoldX, classification was considered correct if the wild-type amino acid was assigned the lowest $\Delta\Delta G$.

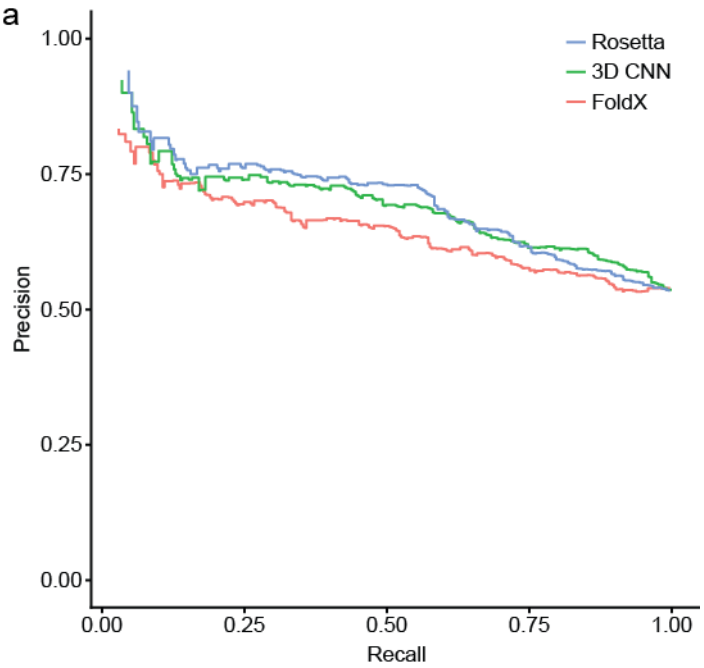


Figure 2.4 Precision recall curve for three computational methods.

Positions where the wild-type residue exhibited the greatest normalized fitness were aggregated from five deep mutational scanning (DMS) datasets. Input data was drawn from DMS datasets and PBD files for the following proteins: TEM-1 β-lactamase, PDB:1BTL; protein G, PBD:2QMT; aminoglycoside-3'-phosphotransferase-IIa, PDB:1ND4; ubiqutin, PDB:4XOF, and Hsp90, PDB:2BRC. The ability of each computational method to identify these positions as wild-type was analyzed as the threshold for classification was varied.

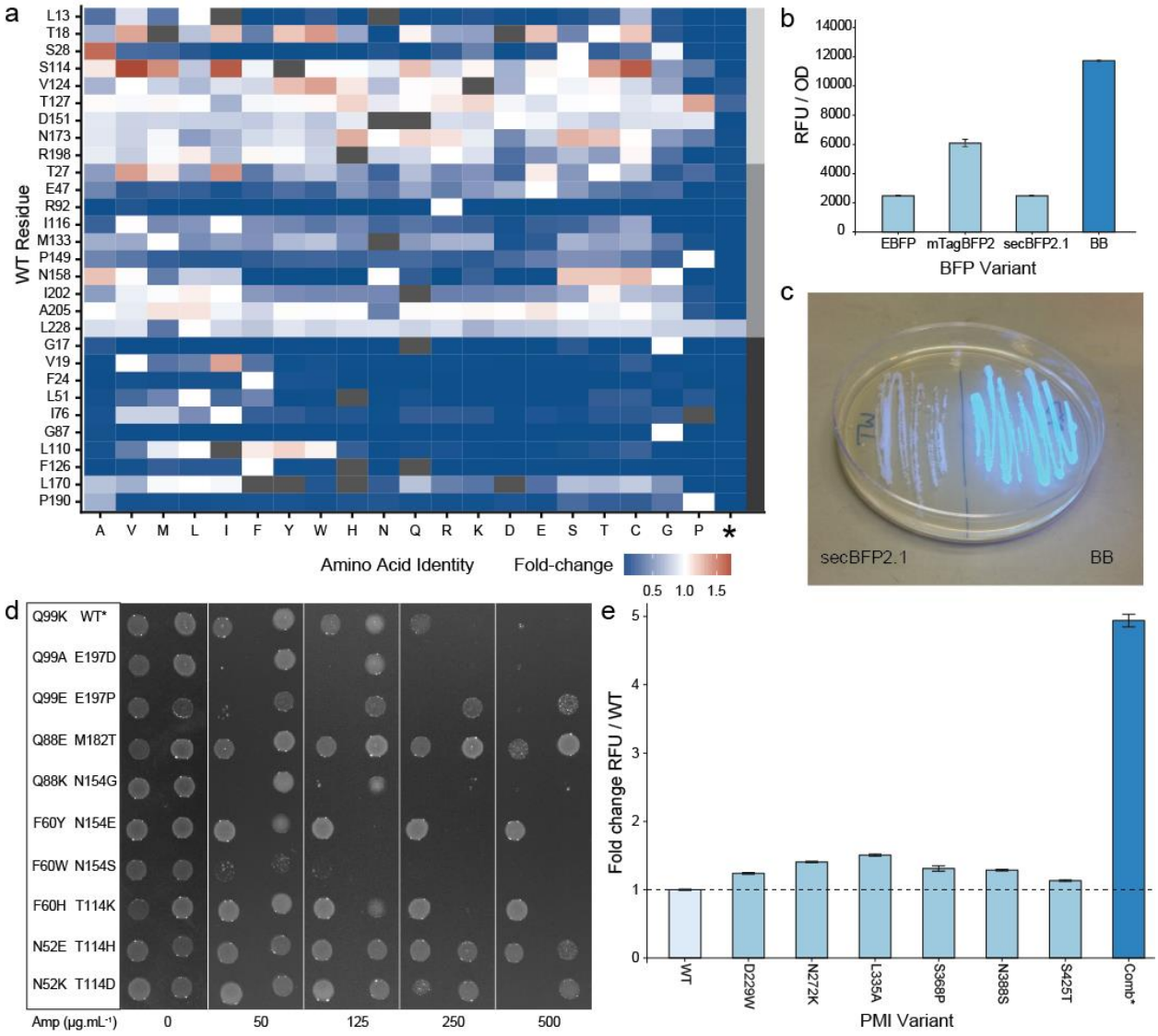


Figure 2.5 Empirical validation of the model as a tool for protein engineering.

a) Heatmap showing fold-change over wild-type for site-saturation mutants of secBFP2.1. The light grey, dark grey and black bars on the right indicate the series of disfavored, random and favored residues respectively. Note, L228 is only five residues away from the C-terminus. Substitutions at this position, including stop codons, have minimal impact on fluorescence. b) An improved variant of secBFP2.1 containing mutations T18W, S28A, S114V, V124T, T127P, D151G, N173T and R198L. was ~6-fold more fluorescent in vivo than the parental protein. This variant was named BFP-Bluebonnet (BB). c) Plate assay showing increased in vivo fluorescence of BFP-Bluebonnet compared to secBFP2.1. d) Stabilizing mutations were identified in TEM-1 β-lactamase at N52, F60, Q88, Q99, T114, M182 and E197. WT* contains the destabilizing mutation L250Q. Residue Q88 was ranked as the 11th least favorable in TEM-1 β-lactamase and was included in place of D214 which lies in the active site. e) Beneficial mutations were identified in CaPMI at residues D229, N272, L335, S368, N388 and S425. A combined mutant containing D229W, N272K, L335A, N388S and S425T was five-fold more fluorescent than wild-type using the split-GFP assay. While S368P was identified as stabilizing by itself, it was deleterious in combination.

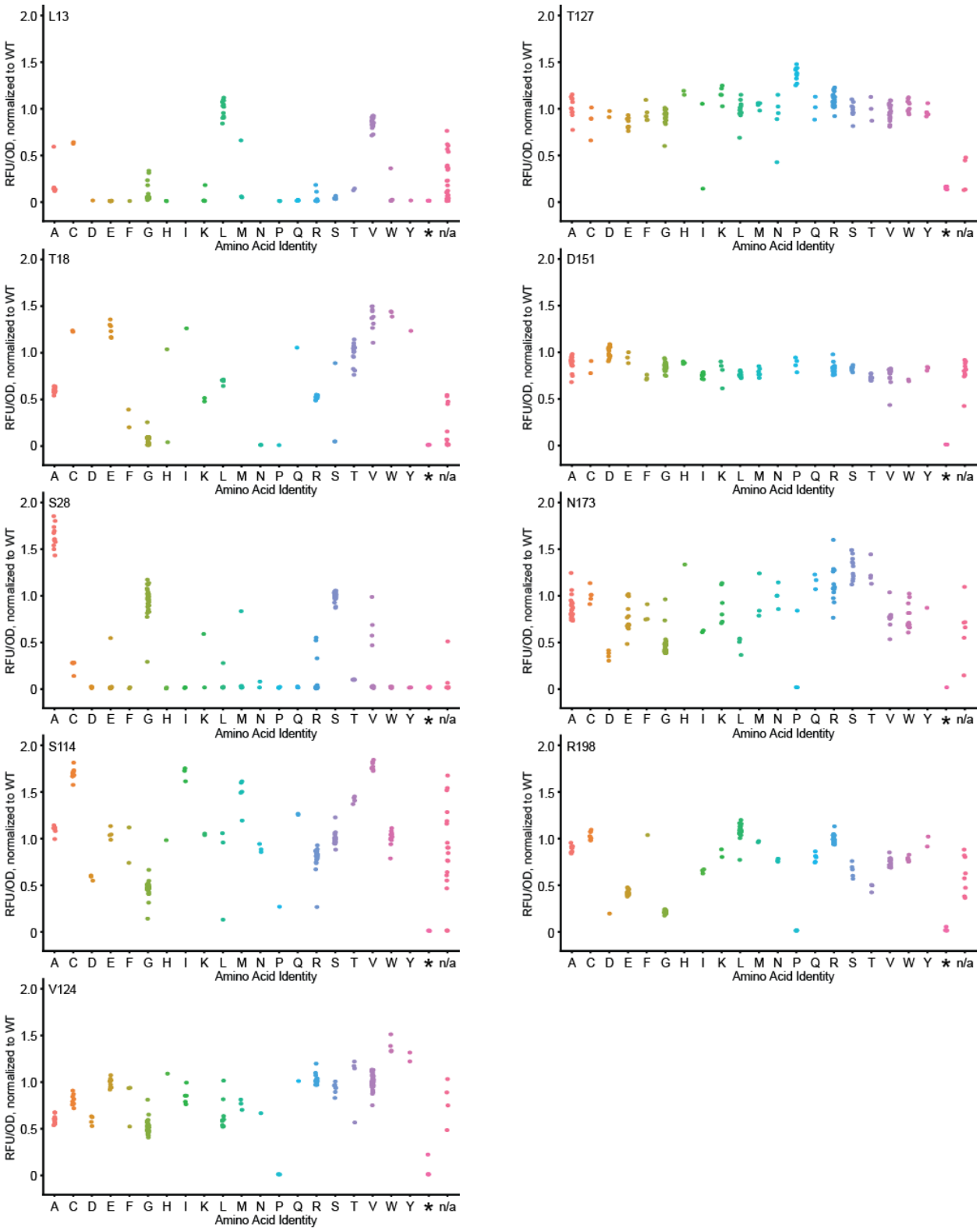


Figure 2.6 Fluorescence data for site-saturation libraries at disfavored residues in secBFP2.1. Raw fluorescence values were normalized to OD_{600} and to the average wild-type value. Outliers were identified through the OPTICS algorithm and removed. n/a represents variant calls that failed to meet the specified thresholds.

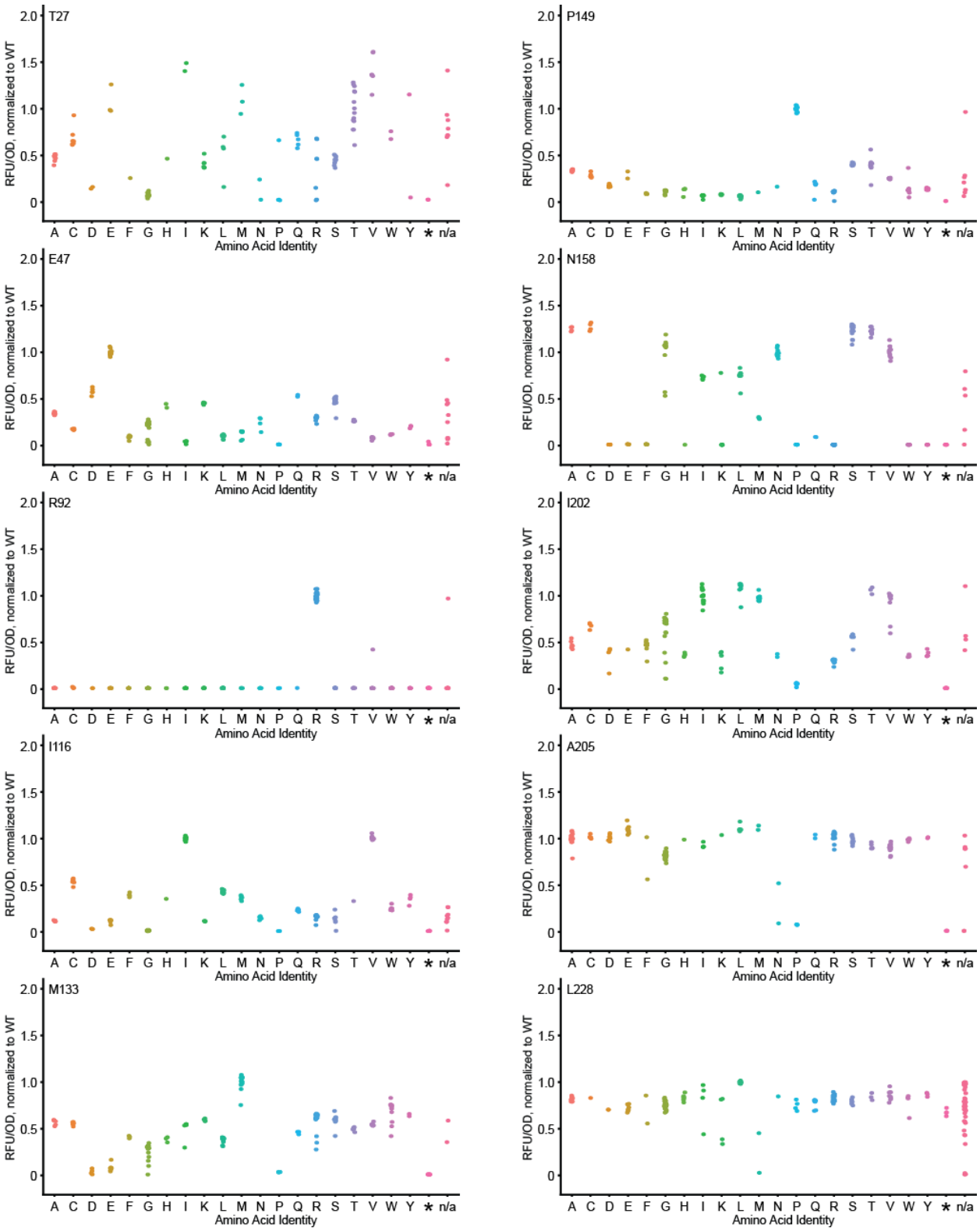


Figure 2.7 Fluorescence data for site-saturation libraries at random locations in secBFP2.1. Raw fluorescence values were normalized to OD_{600} and to the average wild-type value. Outliers were identified through the OPTICS algorithm and removed. n/a represents variant calls that failed to meet the specified thresholds.

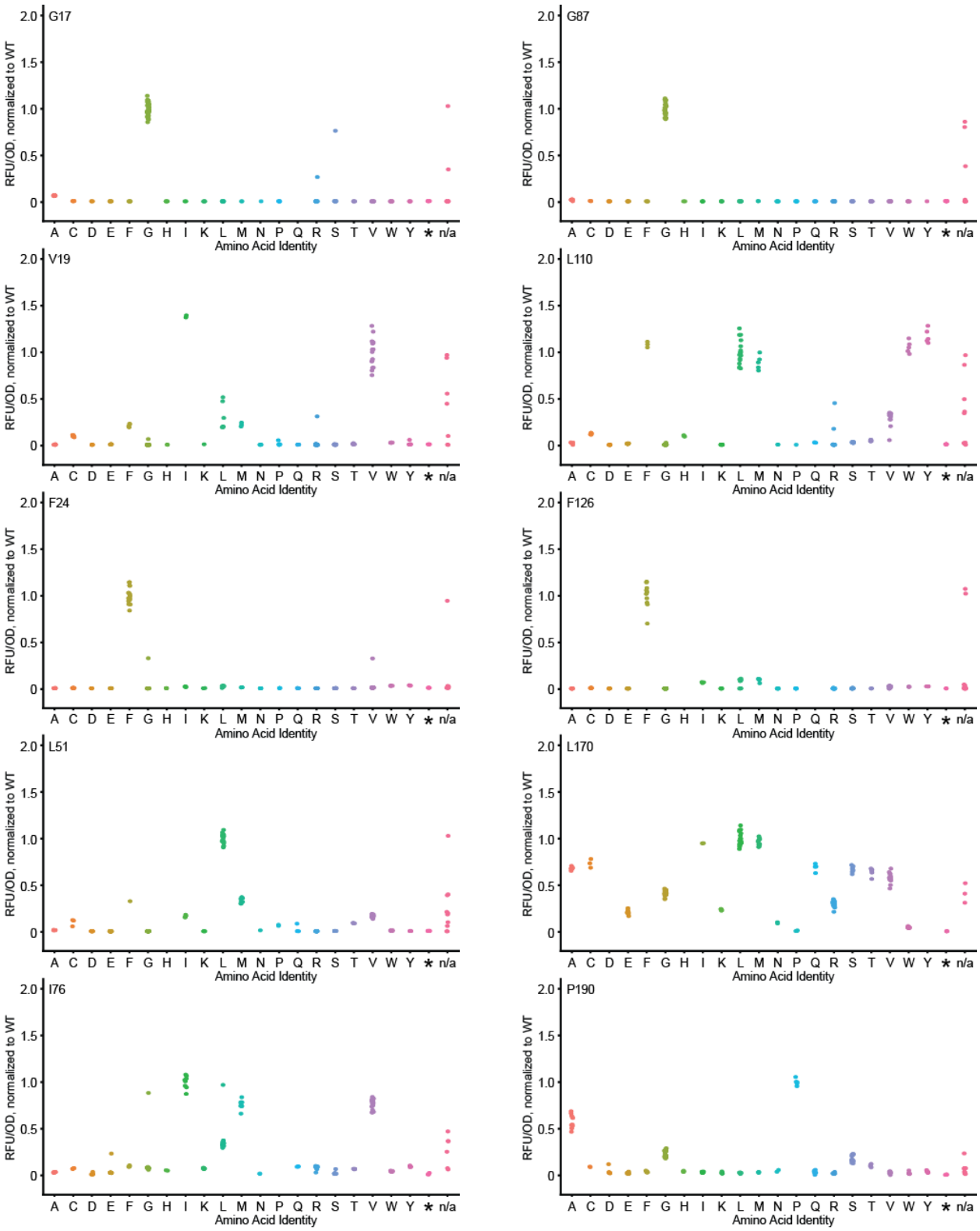


Figure 2.8 Fluorescence data for site-saturation libraries at favored residues in secBFP2.1. Raw fluorescence values were normalized to OD_{600} and to the average wild-type value. Outliers were identified through the OPTICS algorithm and removed. n/a represents variant calls that failed to meet the specified thresholds.

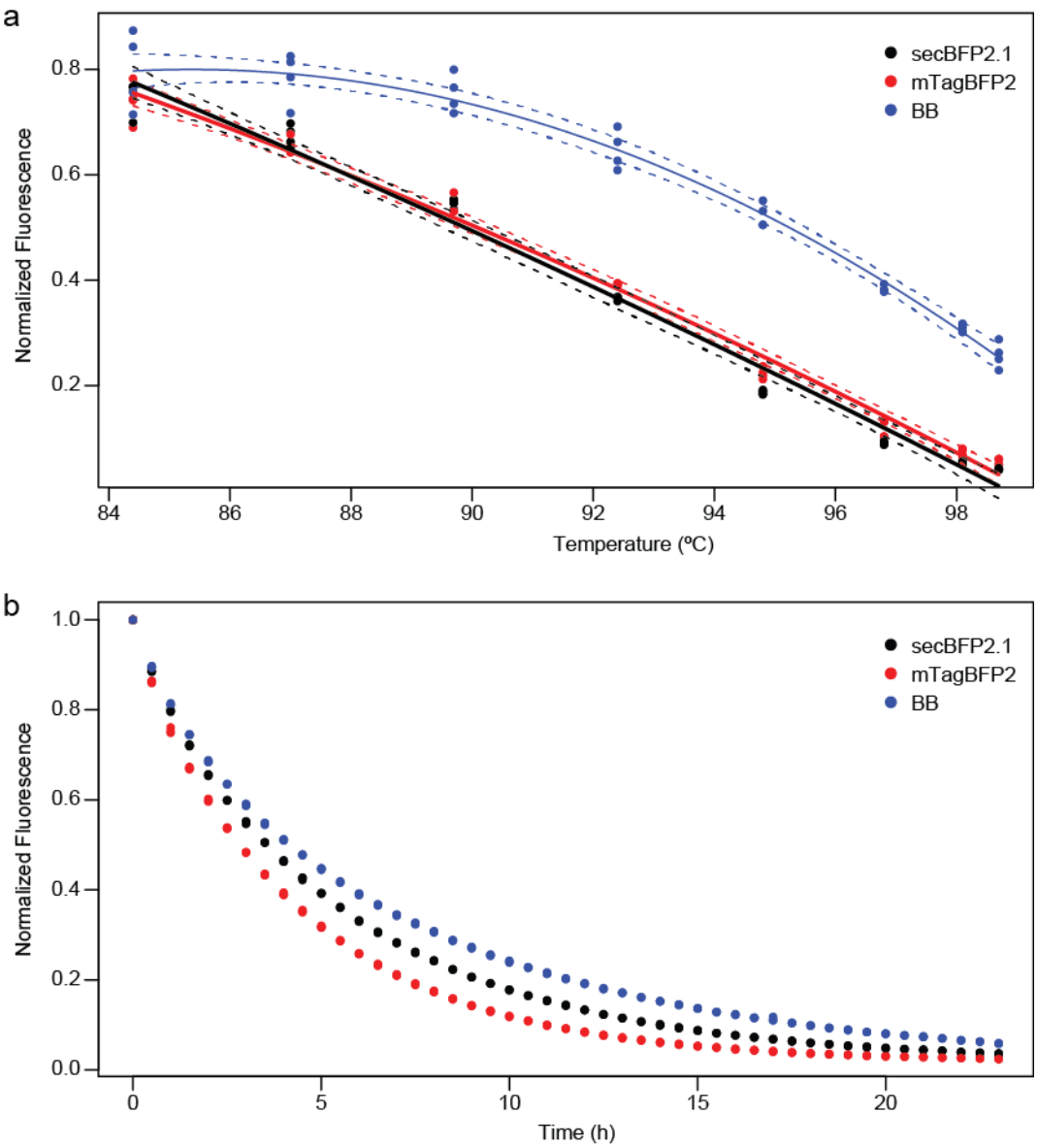


Figure 2.9 BFP-Bluebonnet (BB) exhibited improved folding compared to parental proteins.

a) Plot of residual fluorescence after a ten minute thermal challenge at the indicated temperatures. Quadratic terms were significant in the global linear model and lines correspond to a quadratic model fit for each blue fluorescent protein. Dotted lines delineate 95% confidence intervals for the regression. mTagBFP2 and secBFP2.1 were not significantly different from each other while first order and quadratic terms for BB were significantly different compared to either parental protein. The assay was performed with 4-fold replication. b) Guanidinium melt of BFP variants. This assay was performed with 3-fold replication.

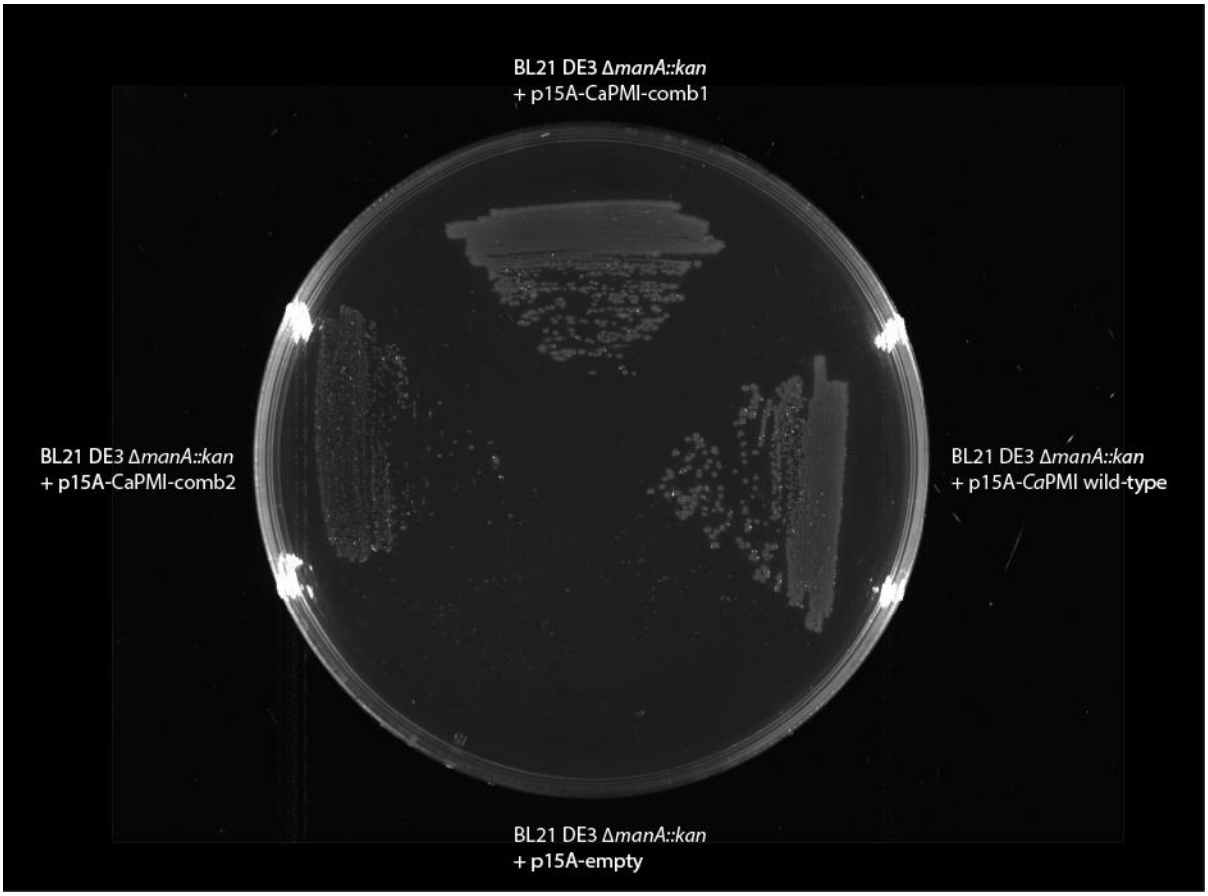


Figure 2.10 Mutant CaPMI variants complement deletion of the E. coli manA gene.

*Ca*PMI-comb1 contains mutations D229W, N272K, L335A, N388S and S425T. *Ca*PMI-comb2 contains mutations S56A, G119A, Q157I, Q193D, D229T, C295V, L335E, K347R, S368N, K402R and Q428T. Growth of *Ca*PMI-comb2 was poorer than wild-type *Ca*PMI and *Ca*PMI-comb1.

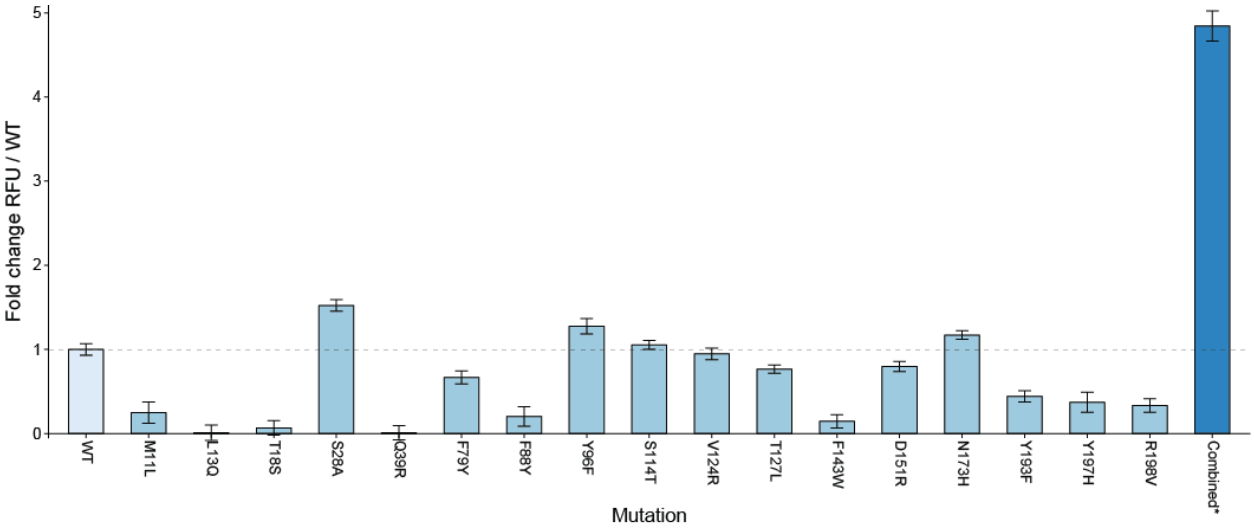


Figure 2.11 Fluorescence assay of secBFP2.1 variants

The combined BFP variant contains mutations S28A, S114T, T127L and N173H. While Y96F was identified as stabilizing by itself, it was deleterious in combination.

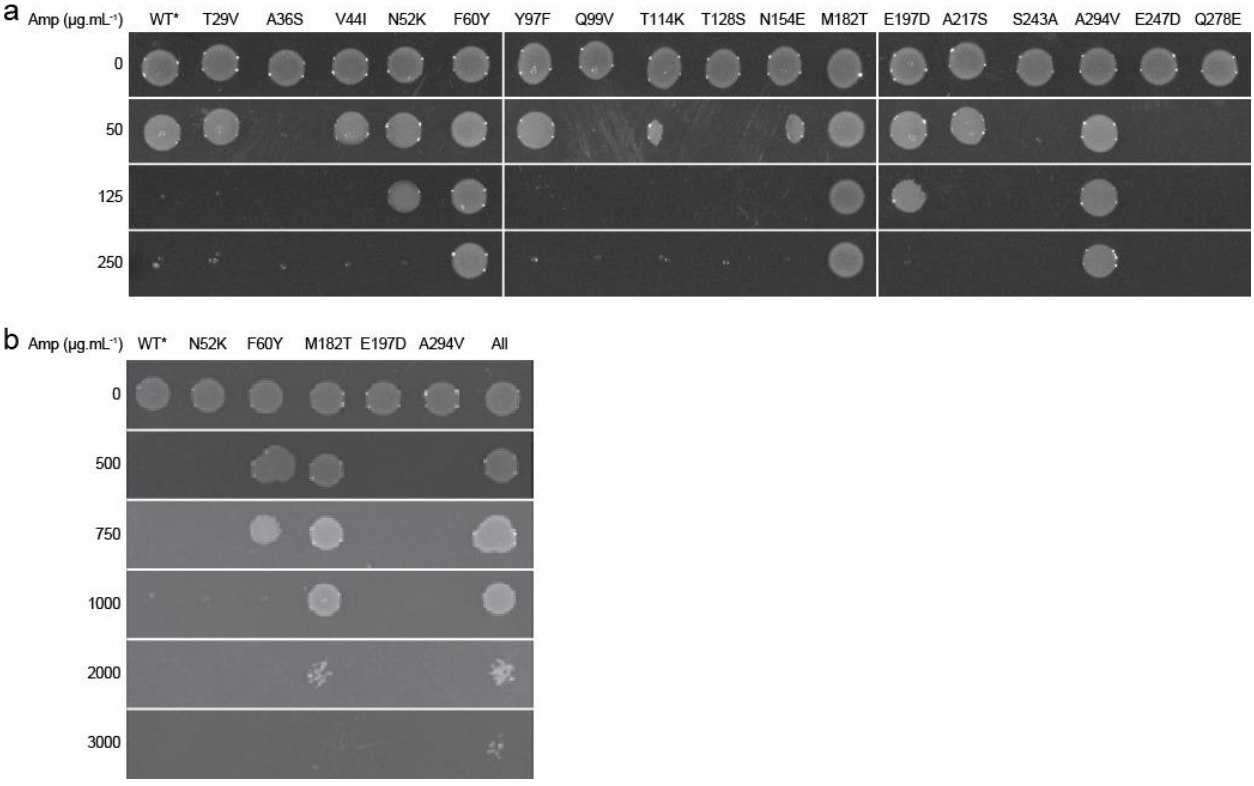


Figure 2.12 Antibiotic resistance assay of TEM-1 β-lactamase variants

Individual mutants N52K, F60Y, M182T, E197D and A294V singularly and a combined variant containing all five stabilizing mutations resulted in increased ampicillin resistance. WT* contains the destabilizing mutation L250Q.

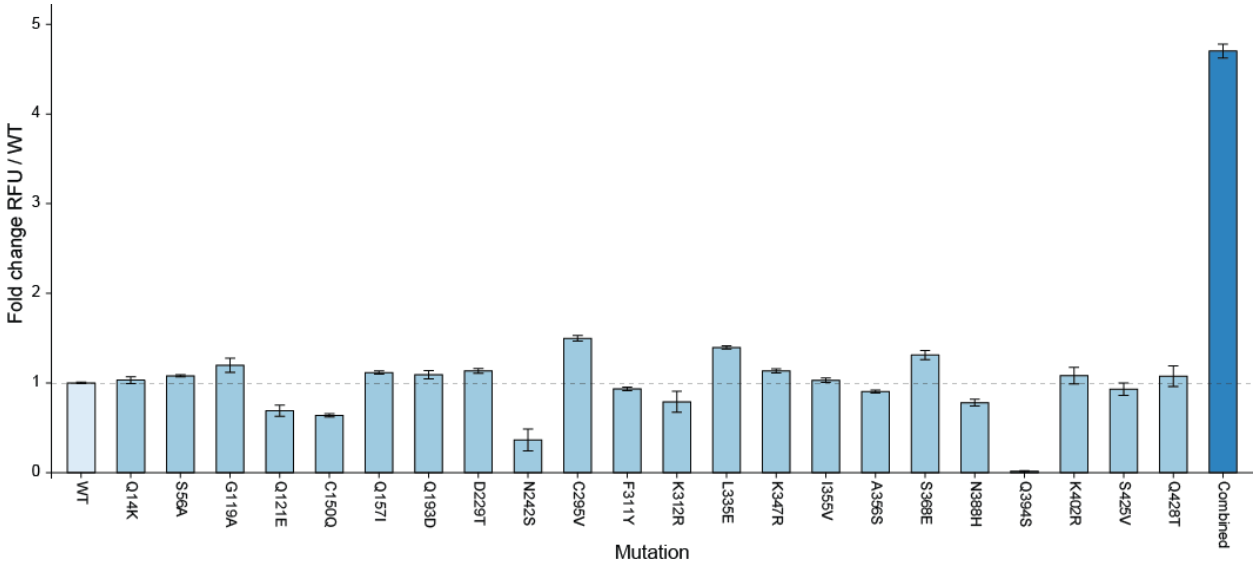


Figure 2.13 Fluorescence assay of the split-GFP-CaPMI fusions.

The combined *Ca*PMI variant contains mutations S56A, G119A, Q157I, Q193D, D229T, C295V, L335E, K347R, S368N, K402R and Q428T.

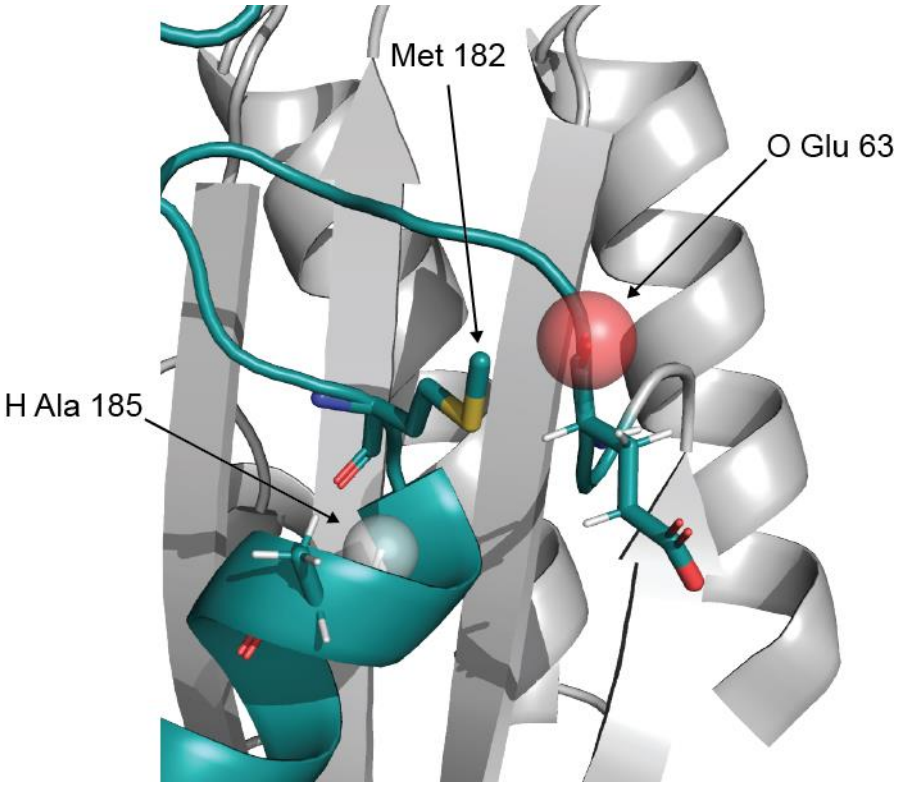


Figure 2.14 Masking of atoms reveals the mechanism of a global stabilizing mutation.

Each atom in the Met 182 microenvironment was systematically deleted and the atoms favoring a mutation to threonine were identified. Of these, the two atoms, O Glu 63 and H Ala 185, change in probability by over 200-fold and have been identified previously in the literature as stabilization pathways for M182T[76].

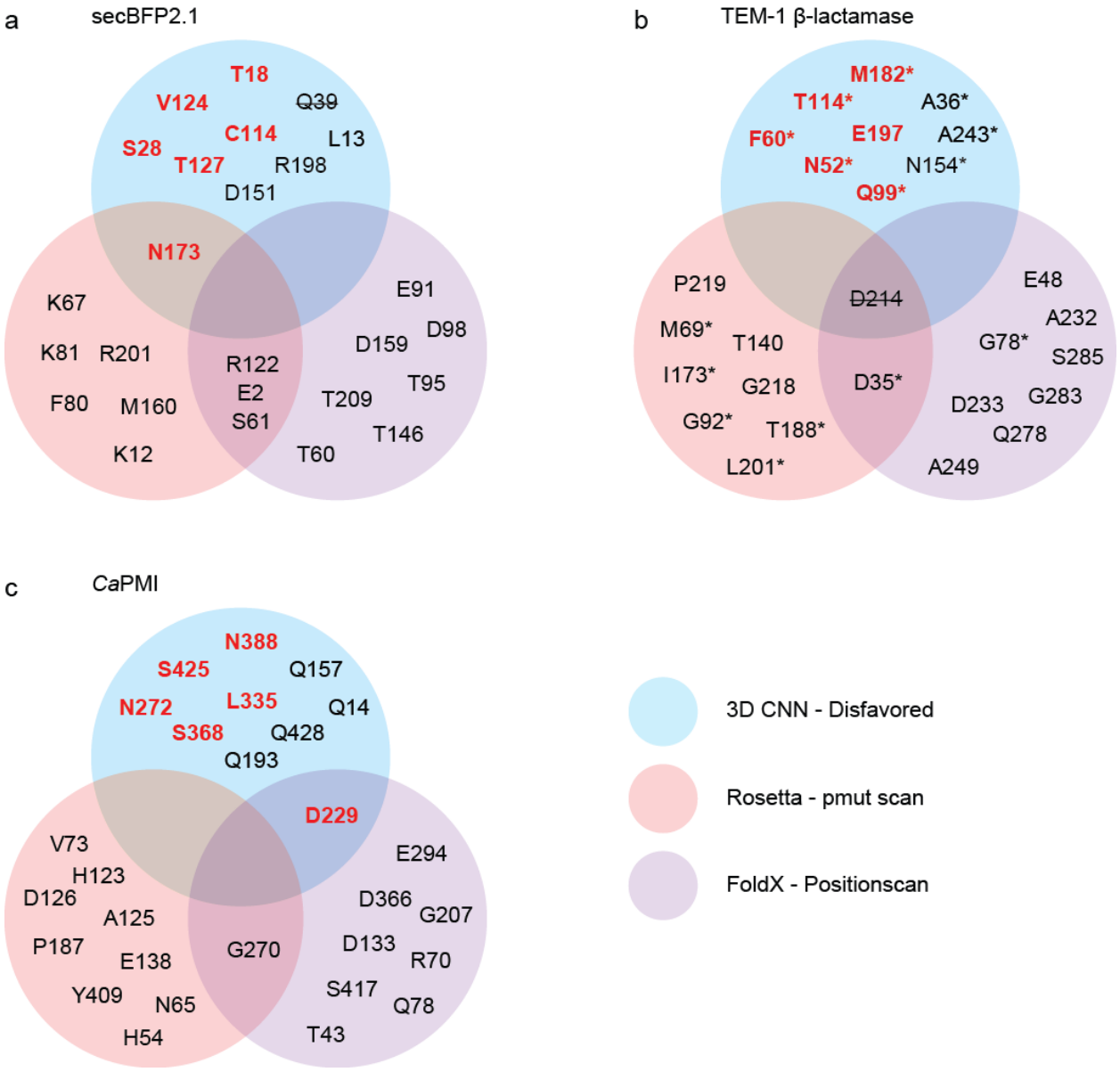


Figure 2.15 Venn diagram showing overlap between different computational tools.

Predictions of the 10 least favorable residues in model proteins made by different computational tools for protein design. Residues colored red indicate positions where we identified beneficial substitutions. Q39 in secBFP2.1 was not analyzed due repeated failure of our site-saturation library to assemble. D214 in TEM-1 β-lactamase was excluded due to its location in the enzyme active site. * Locations in TEM-1 β-lactamase where global suppressors and other beneficial substitutions have been identified[84,85].

Chapter 3

Directed Evolution of Polymerases

Polymerases drive the majority of molecular biology research, and as a result of their ubiquity, are increasingly being tailored to perform more specific functions. This chapter explores how new functionality can be imparted into these remarkable enzymes.

Despite being largely conserved through the tree of life, polymerases can exhibit different properties in different organisms. For example, the majority of DNA polymerases used in molecular biology research are derived from bacteriophages that take advantage of an inherent faster processivity or extremophilic organisms that can withstand the high temperatures needed for polymerase chain reaction (PCR) or tolerate unavoidable inhibitors in a reaction. Yet, these polymerases with slight differences functionally predominantly share the same conserved structural motifs: the 'fingers' position the incoming template, the 'palm' domain contains the catalytic site, and the 'thumb' binds the DNA as it exits the polymerase. The following chapter first asks how novel function from conserved enzymes can be conveyed without requiring knowledge of the functionally critical regions. Put another way, in the context of a DNA polymerase how modular are the domains that give rise to unique functionality?

To this end, we combine the functionality of two family A polymerases—the thermostability of DNA polymerase from *Thermus aquaticus* with the strand-displacing capabilities of the polymerase from *Geobacillus stearothermophilus*. Bst LF and its engineered homologues are the polymerases of choice for high temperature isothermal

---

This chapter is adapted from Milligan, J. N., Shroff, R., Garry, D. J., & Ellington, A. D. (2018). *Biochemistry*, 57(31), 4607-4619., I shared first authorship with JNM.

This chapter was in part adapted from Ellefson, J. W., Gollihar, J., Shroff, R., Shivram, H., Iyer, V. R., & Ellington, A. D. (2016). *Science*, 352(6293), 1590-1593. It is presented with modifications under the full permission of the original publishers.

This chapter was also in part adapted from a draft manuscript: Ellefson, J. W., Shroff, R., Boulgakov, A. A., Hughes, R. A., Marcotte, E. M., & Ellington, A. D. (2019). I shared first authorship with JWE.

My contributions are outlined in the text.

amplification reactions because of their high strand displacement activity[86–89], but they are unstable at temperatures above 70 °C. In contrast, Klentaq has much weaker strand displacement activity but is more thermostable than full-length Taq, with a half-life of 21 min at 97.5 °C[90]. Through molecular breeding of these related but phenotypically different enzymes, we sought to select a chimeric variant that would combine both activities.

Enabled by a directed evolution platform modified to work at mesophilic temperatures (termed HTI-CSR), the work presented here highlights the use of a high-throughput selection to test 10^7 different chimeras to successfully find a polymerase variant that can survive initial denaturation at high temperatures while retaining the strand displacing capability. In this scheme, libraries of polymerase variants are expressed in cells, and cells are ensconced within individual compartments in water-in-oil emulsions. However, genes that produce functional polymerases are not amplified via hyperbranched rolling circle amplification (hbRCA) and enriched for subsequent cycles. Though the end resulted in a useful diagnostic enzyme, the characterization of polymerase variants was performed at low throughput.

Continuing the exploration into polymerases, this chapter next delves into the plasticity of a single DNA polymerase. As an essential cog in cellular integrity, DNA polymerases must exhibit a degree of robustness such that errors are not introduced in the genome and propagated to future generations. Often, this is ensured through a proofreading domain structured in a 3'-5' exonuclease. With this consideration in mind, we chose to begin explore polymerase/substrate interactions with the Archaeal family-B DNA polymerase from the thermophile *Thermococcus kodakarensis* (KOD). Two studies are presented that evolve this polymerase to utilize two different nucleic acid templates. Interestingly, in both examples, the shift from DNA to a new template are accomplished through only a handful of point mutations, suggesting inherent promiscuity within DNA polymerases.

First, KOD DNA polymerase is evolved with the capability to reverse transcribe RNA to DNA. This enzyme, branded as reverse transcription xenopolymerase (RTX) is the first reported thermostable, error-correcting reverse transcriptase, with a demonstrated higher fidelity than other reverse transcriptases. To show this polymerase's use in molecular biology applications, compatibility with RT-qPCR and RNA-seq workflows is presented. The distinct phylogenetic lineage of RTX combined with its research utility suggests even highly conserved machinery with key molecular function are amenable to adding new functionality.

Next, the capabilities of RTX are further expanded by investigating if the evolution towards RNA templates was a unique evolutionary event or a harbinger of usage towards more exotic templates. In a similar manner, RTX is evolved to enable efficient utilization of 2'-OMe templates. The utility of this new polymerase, RTX-Ome, is demonstrated through a novel DNA information encoding scheme where a layer of security is added by storing information by 2'OMe oligos. The directed evolution and encoding/decoding strategy can serve as a platform for custom storage security solutions.

My contributions to this work follow after the directed evolution was performed. After seven rounds of evolution, I assayed the resulting variants to identify candidate polymerases that retained both desired properties defined at the onset of the project. Specifically, I purified polymerase variants from the evolution rounds and assayed for LAMP and RCA activity as well as thermal tolerance. For the subsequent manuscript, I assisted in writing and data presentation. For RTX, my contributions involved taking of advantage of next generation sequencing for a high throughput characterization platform. Traditionally, polymerase fidelity is assayed using plate-based phenotypes, counting bacterial colonies for which an error arising from misincorporation changes an observed phenotype. We wished to develop a higher throughput measure, leveraging the tens of millions sequences provided by NGS. For RTX, this embodied adapting a barcodes to ensure low error reads using this strategy to calculate the fidelity of multiple reverse transcriptases. In demonstrating the utility of RTX-Ome, I developed the bioinformatics pipeline necessary to decode the information from sequencing reads. The process was robust enough to allow for the systematic deletions observed to not corrupt the generated data out. In both resulting manuscripts, I contributed to data analysis and writing.

3.1 Introduction

Over the last several decades, isothermal nucleic acid amplification (IA) has become a transformative technology for point-of-care diagnostics that seek to deliver clinical results to patients in near real time[91,92]. Because IA methods often seek to amplify DNA or RNA via continuous replication at a single temperature, they obviate the need for thermal cyclers and

can reduce the time to result[86,91,93–95]. These assay advantages have in turn enabled the creation of a variety of fascinating and useful point-of-care devices[86,91,92,95–98].

While some IA mechanisms depend upon multiple enzymes, including nickases, recombinases, and ligases, to achieve continuous replication, rolling circle amplification (RCA) and loop-mediated isothermal amplification (LAMP) require only polymerases and primers[87,93,99]. RCA can proceed at mesophilic or higher temperatures, amplifying continuously around a circular template to generate long, concatenated DNA products[100]. When initiated from a nick or single primer, amplification is linear; when both forward and reverse primers are included, however, amplification becomes exponential, generating 10^9-fold amplification in 90 min from as little as 10 copies of template in a reaction commonly called hyperbranched RCA (hbRCA)[88,89,91,99]. LAMP, which is also exponential, is currently an inherently higher-temperature reaction that uses four to six primers to generate 10^9-fold amplification of short (100–500 bp) DNA targets in an hour or less by creating ladder-like concatenated amplicons[87,94,101]. Overall, both methods are rapid, single-enzyme DNA detection systems that are comparable to PCR in terms of sensitivity yet are faster and can operate isothermally, likely explaining their prevalence in point-of-care assays and devices[91,95].

Like many IA strategies, LAMP and RCA rely upon the inherent strand displacement activity of a polymerase to displace downstream DNA, thereby enabling continuous replication without thermal cycling[93,95]. There are only a limited number of polymerases with strong strand displacement characteristics, primarily the large fragment (exo-) of Geobacillus stearothermophilus pol I (Bst LF) for high-temperature reactions (65–70 °C)[102] or the Bacillus subtilis phage phi29 polymerase (ϕ29) for low-temperature reactions (≤34 °C)[103]. These polymerases are also highly processive[102,104], a property that often coincides with strand displacement and that makes them useful for sequencing otherwise difficult DNA molecules[105–107].

Unfortunately, while many IA mechanisms depend upon an initial heating step (~95 °C) for template denaturation[91], ϕ29 and Bst LF are denatured at temperatures at or above 40 and 80 °C, respectively. Thus, some IA reactions require opening reaction tubes and adding polymerase after the heating step, which is both cumbersome and risky because of the common issues of spurious amplification and cross-contamination inherent in ultrasensitive

IA strategies[91,108]. While many IA mechanisms, including LAMP and some versions of RCA, do not necessarily require template denaturation, pre-reaction heating can nonetheless improve the assay sensitivity[109,110], reduce amplification inhibition from crude clinical samples[111,112], and serve as a nucleic acid extraction method for detection of viruses and bacteria[113,114]. Thus, there is a pressing need for thermostable polymerases that possess significant strand displacement activity.

Although Bst LF works at higher temperatures, it is not truly thermostable at temperatures that may be required for DNA denaturation and "hot-start" LAMP. There have been reports of two thermostable polymerases that could potentially be included in a denaturation step in IA reactions such as LAMP: OmniAmp, a viral polymerase from PyroPhage 3173 with DNA polymerase and reverse transcriptase activities[115], and SD Polymerase, a mutant of the well-known Thermus aquaticus (Taq) polymerase[116]. However, neither of these polymerases has been validated for hot-start LAMP, because of either the lack of sufficient thermostability to survive denaturation steps (OmniAmp) or insufficient strand displacement activity for consistent LAMP performance (SD Polymerase).

Directed polymerase evolution by compartmentalized self-replication (CSR) and other methods has previously been used to identify sequence variants of DNA and RNA polymerases that have altered phenotypes such as increased thermostability, incorporation of unnatural or modified bases, reverse transcription, orthogonal promoter recognition, and resistance to enzymatic inhibitors[117–121]. Recently, an isothermal CSR selection was used to evolve the ϕ29 polymerase[122]. While successful, this selection required freezing and thawing cycles for cell lysis to circumnavigate high-temperature reaction steps, which may limit the acquisition of thermostability or other novel phenotypes with enzymes like Bst LF that have higher polymerization temperatures. In addition, the use of random primers resulted in off-target Escherichia coli genome amplification, which can limit selection efficiency. Therefore, we have developed a more robust method, high-temperature isothermal compartmentalized self-replication (HTI-CSR), for engineering thermostable strand-displacing polymerases. HTI-CSR retains the emulsion-based linkage of genotype and phenotype that was established in thermal-cycling CSR but replaces the emulsion PCR step with hbRCA of supercoiled plasmid DNA. This innovation necessitates that a polymerase must have excellent strand displacement activity in order to amplify its encoding gene from the plasmid.

We have demonstrated HTI-CSR by attempting to combine the robust strand displacement capability of Bst LF with the extreme thermostability found in Klentaq. These distantly related enzymes were recoded to ensure maximal overlap, and a shuffled library was created, from which a thermostable chimeric polymerase was selected that enabled one-pot hotstart LAMP. The chimera was also capable of hbRCA amplification from a supercoiled plasmid template, an entirely novel phenotype. Strand displacement seemed to arise from a relatively short Bst substitution into a Klentaq backbone that may alter the conformation of the "thumb" domain common to DNA polymerases. The further development of HTI-CSR should more generally enable the selection of desirable phenotypes in the strand-displacing polymerases typically used for molecular diagnostics and genome amplification.

The molecular basis for life rests on the information flow between DNA, RNA, and proteins[123]. Early notions of a unidirectional central dogma were amended after the discovery of the reverse transcriptase (RT) enzyme[124,125]. The RT family has a single ancient evolutionary origin based on amino acid homology and the presence of RT across multiple domains of life[126]. RTs are involved in processes such as telomere addition, mitochondrial plasmid replication, transposition, and the proliferation of retroviral genomes[127]. It is also hypothesized to be the catalyst in the transition of the RNA to DNA world by providing an avenue to copy RNA into more stable DNA genomes[128].

The progenitor of RT is postulated to be an RNA-dependent RNA polymerase. Because RNA polymerases generally lack an error-checking 3′-5′ exonuclease domain[126,129], proofreading activity is also not present across the RT family, resulting in low-fidelity reverse transcription and characteristic quasispecies behavior in organisms that rely upon it for replication[130]. In contrast to RTs, other DNA polymerase families have evolved exquisite proofreading mechanisms to increase DNA synthesis fidelity during genome replication[131].

Here, we have directly evolved a reverse transcription xenopolymerase (RTX; **Figure 3.7a**) from an error-correcting DNA polymerase using a modified directed evolution strategy[117], reverse transcription–compartmentalized self-replication (RT-CSR) (Figure **3.7b**). RT-CSR enables the simultaneous screening of up to 10^9 polymerase variants for RT activity. We chose the Archaeal family-B DNA polymerases (*polB*) for directed evolution of the RTX as they are monomeric, hyperthermostable, highly processive, and contain

proofreading domains. We initiated evolution using low-stringency RT-CSR (10 RNA residues) with a random library (one or two amino acid mutations per gene) of KOD polymerase variants. As polymerases were enriched, we gradually increased RT-CSR stringency with the stepwise addition of RNA bases into primers. By cycle 18, primers were entirely composed of RNA—requiring reverse transcription of 176 residues to occur every thermal cycle to maintain exponential amplification in the emulsion polymerase chain reaction (PCR).

To explore plasticity, we next evolved RTX in a similar manner as above towards a novel polymerase substrate interaction with the non-standard sugar 2' O-methyl DNA (**Figure 3.12**), which standard polymerases in thermophilic families cannot utilize as a template. Though enzymes that polymerize 2' O-methyl have been engineered[132,133] or derived from viral sources[134], we nevertheless chose this target to show the utility of our directed evolution platform to evolve polymerases with modified DNA backbones. The utility of this novel enzyme is towards encrypted DNA data storage through modified sugar backbones that hinder standard DNA polymerases from reading encoded information. Information is retrieved through an evolved biochemical interaction between a polymerase and a modified DNA substrate. We employ non-canonical sugar analogs that make novel polymerase-substrate pairs, both of which must be known by a receiver to decipher the secured information through replication into standard DNA. By mixing non-standard oligonucleotides containing a privileged, high value message into a population of standard DNA oligonucleotides encoding a deceptive sacrificial cover message, we not only make our hidden message difficult to read, but also conceal its existence (**Figure 3.12a**). Such a scheme may not only be used to secure archival information, but may enable an alternative strategy to covertly transmit information. By using a polymerase with error-correcting capabilities as the starting point for our platform, we ensure high fidelity propagation of encoded information. Apart from acting as a physical impediment, 2' O-methyl DNA is an attractive medium for information storage since it is naturally stable against nucleases, thereby reducing the threat of contamination and integrity loss.

3.2 Results

3.2.1 Variants from directed evolution can perform LAMP

To test polymerases derived from directed evolution with HTI-CSR (**Figure 3.1**), 12 variants from round 7 and six variants from round 5 were cloned into a protein expression vector with an N-terminal 6xHis for initial screening, Of the 18 polymerases we successfully cloned, five variants could not be expressed and purified in sufficient quantities for screening; these largely consisted of chimeric proteins that had a Klentaq backbone with a short, C-terminal Bst segment, suggesting that this protein configuration may not be compatible with the expression vector, E. coli strain, or purification protocol we utilized.

We initially screened the 13 purified variants for thermostability and strand displacement activity using loop-mediated isothermal amplification (LAMP) with a well-known GAPDH template as an assay. When monitored on a real-time instrument, these reactions yield readily interpretable qPCR-like exponential amplification curves that are widely regarded in the isothermal amplification field as the preferred reaction monitoring method over end-point measurements such as gel electrophoresis[86,87,94,95,98,101]. These real-time LAMP reactions were assembled and monitored on a LightCycler 96 qPCR machine as described previously[98,101], except that we used EvaGreen intercalating fluorescent dye rather than oligonucleotide probes. Amplicon homogeneity was monitored via postreaction melt curves, a strong indicator of amplicon specificity[109].

Initial screening identified two highly functional variants capable of LAMP, variant 5.9 from round 5 (v5.9) and variant 7.16 from round 7 (v7.16) (**Figure 3.2a**). Cq values from the software, which represent the time at which fluorescence exceeds a determined threshold value, indicated that v5.9 was 24.7 min slower (66.56 min) than purified wild-type Bst LF (41.84 min), while v7.16 was 7.3 min faster (34.56 min). Thus, v7.16 has a higher polymerization rate in LAMP reactions than wild-type Bst LF, while v5.9 is slower. In addition, both polymerases were capable of RCA (**Figure 3.3**). We also attempted to compare v5.9 and v7.16 with SD Pol, a commercially available Taq mutant reportedly capable of LAMP[116]; however, we found that in our hands this enzyme was unable to amplify products via LAMP under the recommended conditions, though it was capable of RCA from a nicked plasmid template (**Figure 3.4**). Because they were capable of generating significant

amplification in multiple isothermal reaction formats, we chose to further investigate the properties of v5.9 and v7.16.

3.2.2 Thermoresistance is enhanced in the evolved polymerase variant, v5.9

One key advantage of a more thermostable strand-displacing polymerase would be the ability to carry out isothermal amplification reactions that rely on strand separation, such as LAMP, without the need to add polymerase after the initial template denaturation step. To that end, "one pot" reactions were set up with v5.9 and v7.16 in which the entire reaction mixture, including polymerase, was preheated for 1 min at 85, 89.5, or 95 °C prior to carrying out the remainder of the real-time LAMP amplification reaction. Remarkably, v5.9 successfully performed LAMP after heating at all of the tested temperatures, while neither Bst LF nor v7.16 were able to amplify after any of the thermal challenges (**Figure 3.2b**). Variant 5.9 performed similarly after heating at 85 or 89.5 °C as in assays without preheating of the enzyme (**Figure 3.2c**). Interestingly, the threshold time was slightly shorter after heating at 89.5 °C than at 85 °C, possibly as a result of high temperature denaturation of an enzymatic inhibitor that carried over from purification. Variability within reactions with heat treated enzyme was notably higher; this is likely due to small variations in the heating procedure from run to run, as reactions had to be transferred by hand from the thermocycler to the qPCR machine in order to utilize a temperature gradient for heating. Thus, v5.9 acquired novel functionality not seen in either parent enzyme used to generate the library: both the thermostability and strand displacement activity needed to perform one-pot LAMP.

In order to further characterize the thermal tolerance of variant 5.9, we performed activity assays similar to radioactivity based assays previously described[90,135,136]. We determined the activity by measuring initial reaction rates after heating at 85, 89.5, or 92.5 °C for 1–10 min and normalizing these to the activity without heating (**Figure 3.5**). This assay has previously been used to characterize the "thermal inactivation", or thermoresistance, of Klentaq and Taq[90]. While not a direct measure of thermostability, this method is ideal for characterizing polymerase functionality within the context of diagnostic amplification reactions by measuring the activity after high temperature incubation steps that are typical of such reactions. Variant 5.9 retained 75% activity after heating at 85 °C for 5 min, with a half-life of approximately 6.5 min. At 89.5 °C, the enzyme had a half-life of

approximately 3.2 min, while at 92.5 °C, the half-life was ~1.5 min. This represents significantly less thermal tolerance than Klentaq, which has a half-life of ~21 min at 97.5 °C and shares 97.5% protein sequence identity with v5.9.42 However, it is still significantly more thermostable than Bst LF, which loses all activity within 1 min at 85 °C (**Figure 3.2**). Thus, v5.9 is sufficiently thermotolerant to withstand short incubations at temperatures as high as 92.5 °C, such as those used for pre-reaction template denaturation in isothermal amplification applications.

In the original crystal structure characterizations of Bst LF and Klentaq[102,137], it was noted that increased ratios of certain amino acids (E:D, L:I, and R:K) and increased numbers of prolines can be indicative of higher thermal tolerance. Thus, it is notable that the mutations we observe in v5.9 shift most of these ratios toward decreased thermostability compared with Klentaq: three leucines were lost compared with only one isoleucine, three lysines were gained compared with a loss of one arginine, and two prolines were lost. This correlates with the observed loss of thermostability of v5.9 relative to Klentaq, though the relationship may not be causal.

3.2.3 v5.9 has novel hyperbranched RCA activity

The combination of thermostability and strand displacement characteristics found in v5.9 may prove useful for other reactions as well. We therefore tested v5.9 against its parent enzymes in high-temperature hbRCA reactions[88]. Reactions were monitored via EvaGreen incorporation on a LightCycler 96 qPCR machine. Real-time RCA analysis is much more sensitive and informative than end-point gel electrophoresis, which often requires the interpretation of concatemer smears[138]. In contrast, the real-time amplification plots allow us to measure the polymerization rate and product yield by monitoring the fluorescence signal and differentiate unique strand displacement characteristics on the basis of the observed enzyme kinetics. The use of an intercalating dye on a real-time quantitative PCR machine also enables melt curve analysis of reaction products, which can be used to characterize amplification specificity[109].

While some hbRCA reaction formats utilize relaxed circular DNA templates with multiple forward and reverse primers to enable exponential amplification[100], those that

function with supercoiled templates typically utilize ϕ29 polymerase, known for its extreme processivity[99,103,104,139,140]. In order to effectively replicate within our HTI-CSR selections, however, polymerase variants must be able to replicate from supercoiled plasmids. In order to characterize this activity, we compared the polymerase activities of v5.9, Bst LF, and Klentaq using both supercoiled plasmid templates and plasmids that were nicked in order to relax their supercoiling. All three polymerases were able to replicate DNA using hbRCA with the relaxed template, producing similar amounts of product (**Figure 3.6a**). The reaction rates were similar for v5.9 and Bst LF using the relaxed template, both of which were 2- to 3-fold faster than Klentaq (**Figure 3.6b**). This is likely due to increased strand displacement activity as observed in the LAMP reactions above.

The results were dissimilar when supercoiled plasmids were used. Only v5.9 was able to replicate effectively (**Figure 3.6c**), with a yield comparable to that of the reaction containing the relaxed template. Additionally, v5.9 had a much higher reaction rate than the other enzymes under these conditions (**Figure 3.6d**), although all of the rates were significantly reduced compared with the relaxed template reactions. These data demonstrate that v5.9 has acquired the unique ability to replicate from supercoiled plasmid DNA via hyperbranched RCA, an activity that is not present in either of the polymerases used to generate the shuffled library.

3.2.4 RTX proofreads on RNA and DNA templates

To understand how our process reshaped KOD polymerase to use RNA templates, we deep-sequenced RT-CSR cycles to recapitulate the evolutionary path to RT activity (**Table 3.1**). Mutations were identified throughout the polymerase and accumulated along the template-binding interface so as to progressively increase the length of RNA that could be accommodated. The mutated positions are hypothesized to be molecular checkpoints used to enforce strict DNA template utilization: as the template enters, near the active site, and at the nascent duplex.

Given that RTX is capable of proofreading during reverse transcription, we hypothesized that it may have increased RT fidelity compared to natural polymerases. Barcoded primers used during RT of several human mRNAs allowed multiple reads of a single cDNA during deep sequencing—reducing background sequencing errors by several orders of

magnitude[141] (**Figure 3.10**). Sequencing analyses revealed that the control retroviral RT [Moloney murine leukemia virus (MMLV)] had an error rate of 1.1×10^{-4} to 4.8×10^{-4}, whereas RTX had an error rate of 3.5×10^{-5} to 3.7×10^{-5} (3- to 10-fold lower) (**Figure 3.9b**). The mutational spectra of RTX favored G-to-A transitions and G-to-T transversions, which accounted for nearly half the observed mutations. Inactivating the RTX's proofreading capabilities increased error frequency nearly threefold, supporting evidence that active proofreading was occurring during RT. Inactivating the proofreading of RTX shifted the mutational bias (**Figure 3.9b** and **Table 3.2**). Given that the barcoding error detection limit is identical to the observed error of RTX[141] (**Table 3.2**), we anticipate the true error rate for RTX to be even lower than reported.

3.2.5 RTX can streamline established RNA workflows

RTX has the potential to streamline workflows (combining RT and PCR steps) and increase the precision of transcriptomics, reducing biases and errors introduced in the reverse transcription step of RNA-sequencing protocols[142]. To demonstrate its utility, we introduced RTX into a commonly used platform for RNA sequencing. Analysis revealed nearly identical coverage and expression profiles (**Figure 3.11**), suggesting that RTX is compatible with established workflows. Accurate quantification of input RNA is achievable with RTX using a one-step qRT-PCR of Zika virus-derived RNA templates in a one-pot reaction (**Figure 3.8**). Despite the fact that a TaqMan probe was not involved, negligible non-specific signal was observed with either of these approaches. Moreover, target-derived amplicons could be readily identified by their characteristic melting temperatures.

3.2.6 RTX can be further evolved towards xDNAs

Following RT-CSR process by increasing the number of challenge bases to 81 2' O-methyl bases at the final and 18th round, we used next-generation sequence to probe high frequency mutations after the final round of selection. Using the mutational information provided by next generation sequencing of the final library round (**Table 3.3**), we constructed a series of variants, composed of a high frequency of observed mutations and that were unlikely to inactivate the proofreading domain .

We next wished to demonstrate the utility of RTX-Ome in a cryptogenetic approach. In particular, we sought to store and recover privileged information, accessible only with a paired physical key. Using the previously reported encoding scheme DNA Fountain[143], we transformed a series of innocuous text files into unmodified DNA to act as a sacrificial cover message (**Figure 3.13a**). In a similar manner, files apropos to information security and cryptography like the Zimmerman Telegram, Cryptographie Indechiffrable, and the Kryptos Panels among others were encoded into 2'Ome DNA (**Figure 3.13a**). In total, the cover and hidden files were encoded into 4000 and 2000 oligos, respectively, with only the hidden oligos containing a modulo 2000 seed to ease downstream recovery. Oligonucleotide pools were individually synthesized on a 12k Customarray oligonucleotide chip, with a 16 nucleotide seed region for positional identification, 64 nucleotides of data containing payload, and 8 nucleotides containing a Reed-Solomon code. As redundancy is built in into the encoding scheme, our simulations revealed that we required an average of 2784 +/- 58 oligonucleotides to recover the cover message and 1245 +/- 46 sequenced oligonucleotides to decode the hidden file (**Figure 3.14**).

Our next goal was to show informatic recovery of the secure oligonucleotides. To accomplish this, we devised an experiment where the standard DNA and 2'Omethyl DNA were mixed in a 1:20 ratio. The pooled oligos were then subjected to an RT-PCR program to both amplify the nucleic acids, as well as append appropriate adapters for Illumina sequencing. Four polymerases were assayed for their capabilities in recovering both the standard and hidden oligonucleotides. In addition to the evolved RTX-Ome, we tested KOD and RTX, the parental origins of our evolved polymerase as well as the two enzyme mix of MMLV/Taq to serve as a positive control.

Unexpectedly, following sequencing of our oligonucleotide libraries we discovered deletions in virtually every sequencing read. Because deletions were systematic across all libraries, we attribute these errors arising from oligo synthesis (**Figure 3.15**) DNA decoding schemes currently do have error-checking mechanisms like the use of a Reed-Solomon code, yet these are only suitable for correcting substitutions and not indels where the majority of oligonucleotide synthesis errors occur. We sought to reconstruct the missing bases by assuming deletions appear randomly thus redundancy in oligonucleotide synthesis would yield intact regions where others are missing. We created bins of similar oligonucleotides

through sequence clustering and performed multiple sequence alignment within each bin to build a consensus sequence. If the length of the consensus sequence was less than the desired length, gaps were filled by inserting positions at which a non-gapped base occurs most frequently and iterated through to find a sequence that matched the designed GC content, homopolymer stretch, and Reed-Solomon code. This strategy generated a consensus sequence for each clustered bin (**Figure 3.16**).

We then used DNA Fountain to decode the messages from our consensus sequences. Due to the widespread deletion errors, we utilized the aggressive decoding flag where input sequences are shuffled and run for many trials. We modified the native functionality slightly by only randomizing sequences that required more than 20 iterations to find a consensus sequence matching our given parameters rather than the full consensus sequence list. RTX-Ome, along with the other three control polymerases, successfully amplified the standard oligonucleotide sequences and decoded the cover files. In decoding the hidden message, MMLV-Taq and RTX-Ome, the only of the three Archeal family B polymerases, were able to correctly recover the secured files (**Figure 3.13b**). In cases where trials produced different checksums as with KOD and RTX-Ome, the correct checksum was observed most frequently and no other checksum appeared more than once (**Table 3.4**). To show robustness of our decoding strategy, we performed random 10% down sampling of the total sequencing reads and observed fully correct recovery as with the full reads. These results highlight the ability to recover secure information with a xDNA/polymerase physical encryption.

3.3 Discussion

Compartmentalized self-replication[117] typically relies on multiple thermal cycling steps and has been used primarily with thermophilic polymerases to evolve novel functionalities that can improve PCR. We now describe how a variant of CSR (HTI-CSR) can be adapted to the directed evolution of strand-displacing polymerases for isothermal amplification reactions. HTI-CSR is notable for being an emulsion-based selection that can allow large libraries to be sieved but requires only a single thermal lysis step to accommodate both thermostable and mesothermophilic polymerases. Recently, Povilaitis and co-workers developed a similar isothermal, emulsion-based directed evolution scheme that relied on

whole genome amplification (WGA) at low temperatures to evolve a ϕ29 polymerase mutant with improved thermostability (up to 42 °C) and increased amplification rate[122]. In contrast, our HTI-CSR selection requires only a single heating step to accomplish lysis, leaving cells intact prior to compartmentalization. Moreover, the use of specific primers in our thermostable selections provided added specificity, and the reduction of the number of primers used in a given round led to an increased stringency of selection, as the polymerase had to reproduce its gene via ever longer extensions.

We applied HTI-CSR to a library of chimeras generated by shuffling the Family A polymerases Bst LF and Klentaq. Bst LF is a well-known strand-displacing enzyme that is often used in isothermal amplification reactions, while Klentaq is very thermostable and is a mainstay for PCR; it was hoped some sequence combination of the two would yield an enzyme with both properties.

Selection of the overlap library did indeed lead to enzymes with a variety of phenotypes. For example, sequencing data suggested that our polymerase library had become errorprone by round 7 and as a consequence accumulated polymerases like v7.16 that survived the selection despite lacking thermostability, likely as a result of an increased replication rate at the expense of fidelity. This phenomenon has been observed in other CSR selections[119,121]. In contrast, v5.9 did not contain mutations known to decrease fidelity and proved to be a thermostable polymerase with greatly improved strand displacement capabilities. Interestingly, modeling of v5.9 based on the structures of its two parent enzymes yielded an interesting new insight into the poorly understood process of strand displacement. While Klentaq is not normally a strand-displacing enzyme, a small Bst LF substitution may lead to changes in the secondary structure at the base of Klentaq's thumb subdomain that in turn fully enable a versatile strand displacement phenotype.

The v5.9 chimera proved to be useful for carrying out hotstart LAMP and a variety of high-temperature RCA reactions. The enzyme survived 1−2 min template denaturation steps at ∼90 °C, which were sufficient for complete template denaturation in our LAMP assays. The ability to subject preassembled isothermal amplification reactions to high temperature incubations will certainly be useful for diagnostic applications, as pre-reaction heating can improve assay sensitivity[109,110], reduce amplification inhibition from crude clinical samples[111,112], and serve as a nucleic acid extraction method for the direct detection of

viruses and bacteria[113,114]. This is in stark contrast to typical isothermal amplification protocols, which require opening tubes to add enzyme following thermal denaturation to protect the enzyme, creating an unwieldy and complicated workflow. Such "one pot" reactions may be especially useful for point-of-care applications, where the high-temperature step could now be included in lab on a chip devices[144–149]. Furthermore, v5.9 performed similarly or better than Bst LF in all of the isothermal mechanisms we tested; its superior performance in hyperbranched RCA from supercoiled templates may make v5.9 preferable for many applications.

In order to provide a preliminary structural context for the observed functional properties of variant 5.9, we compared the crystal structures of its ancestors, the large fragments of Bst and Taq polymerases (Bst LF and Klentaq)[150,151]. When the Bst LF and Klentaq structures are aligned, we see that the structures are quite homologous, especially in the finger and thumb subdomains. Sanger sequencing had revealed that variant 5.9 is a chimera consisting mostly of Klentaq (97.5% sequence identity) with a small 14 amino acid chimeric section identical to Bst LF and two nonsynonymous mutations, E322G and L484S. We mapped this chimeric region of v5.9 onto the structures of Klentaq and Bst LF for comparison. The inserted region was located in the polymerase domain at the base of the thumb subdomain, just below the I helix, an essential structure of the thumb subdomain, forming an antiparallel coiled-coil structure with the neighboring H helix that is dependent on hydrophobic interactions between leucine residues on the two helices[102]. We hypothesize that the Bst LF substitution and additional mutations observed in v5.9 may be stabilizing a particular conformer of the thumb subdomain that is important for the observed strand displacement characteristics.

Before v5.9, the only polymerase known to be capable of hyperbranched RCA from supercoiled plasmids was ϕ29, which amplifies DNA ~160-fold in 6 h using unmodified random hexamer primers[99]. Others have shown that gene-specific primers can improve hyperbranched RCA amplification as much as 15-fold with ϕ29[149], suggesting that ϕ29 is capable of roughly 2400-fold amplification in 6 h using assay conditions similar to ours. This is comparable to the 1500-fold amplification in 4 h we observed with v5.9 in our fluorescence assay, suggesting that these polymerases have similar hyperbranched RCA activities. In view

of this, v5.9 may also be useful for similar applications, such as rapid amplification of plasmid DNA for sequencing applications[99].

HTI-CSR may now allow the directed evolution of many different polymerase phenotypes, including altered nucleotide specificity, resistance to inhibitors, and the utilization of new templates. Our initial selection optimization with wildtype Bst LF also suggests that lysozyme-mediated HTI-CSR could be used to optimize multi-enzyme reactions, such as the polymerase and nickase together for RCA. Such co-optimizations might greatly improve isothermal amplification reactions for diagnostics, where efficiency can often be stymied by something as simple as a dissonance in buffer conditions between enzymes[95,138].

Using RT-CSR, we have altered the substrate specificity of a high-fidelity DNA polymerase, highlighting the plasticity of highly conserved molecular machinery. Only a handful of mutations were required to impart RT activity, suggesting that the evolutionary hurdle for forming high-fidelity reverse transcription is relatively low. Nevertheless, all known retroelements use proofreading-deficient RTs, suggesting that high error rates are either a historical coincidence or an evolutionary strategy to promote diversity. Another possible explanation is that high fidelity was never required simply because RNA genomes are small as a result of their inherent instability[152]. Given the plasticity of these polymerases for modified templates and the adaptability of the RT-CSR framework (as primers are simply programmed to contain modified bases), RTX evolution should be compatible with many base and sugar analogs[153–156]. Combination with previously evolved XNA polymerases could enable synthesis of genomes entirely composed of artificial nucleic acids[157].

The xDNAs have a number of advantages over storage of information in DNA. For instance, most xDNAs have remarkable stability towards nucleic acid degrading enzymes due to their non-natural chemical structure. This minimizes the need for sterile conditions in future DNA storage centers which should greatly reduce costs. Perhaps most importantly, xDNA provides a high layer of security due to having to know the chemical composition of the DNA and the significant effort and expertise required to generate polymerases which are capable of reverse transcribing the xDNA. This platform benefits from using error-correcting polymerases, which when dealing with encoded information, ensures faithful replication of data.

3.4 Methods

3.4.1 Polymerase purification

Wild-type Bst LF and Klentaq as well as individual variants isolated from selection were cloned into pATetO 6xHis (see Strains, Primers, Plasmids, and Cloning). Products were amplified using primers JNM316 and JNM309 for the Bst LF 5′ and 3′ ends and primers JNM317 and JNM310 for the Klentaq 5′ and 3′ ends, respectively. Plasmids were transformed into BL21 cells. Single colonies were inoculated into 5 mL of Superior Broth (Athena Enzyme Systems) supplemented with 100 μg/mL ampicillin and grown overnight at 30–37 °C. Cultures were diluted 1:200 into 250 mL to 1 L of fresh medium, cultured to OD600 0.5–1, induced with a final concentration of 200 ng/mL ATc, and further cultured for 3–7 h for expression. Cells were harvested (4000g, 15 min, 4 °C), frozen in liquid nitrogen, and stored at −80 °C. Cells were resuspended in 20–40 mL of lysis buffer (20 mM Tris, pH 7.4, 300 mM NaCl, 0.1% Tween-20 (Thermo Fisher Scientific), 10 mM imidazole) supplemented with EDTA-free Protease Inhibitor Tablets (Thermo Fisher Scientific) and 0.5 mg/mL lysozyme and mixed end-over-end for 30 min at 4 °C. Cells were further lysed using sonication. Supernatants were cleared (40000g, 30 min, 4 °C), heated for 65 °C for 20 min with shaking (400 rpm), and cleared again (20000g, 20 min, 4 °C). Polymerases were purified by metal ion chromatography. Briefly, lysates were added to 1 mL of preequilibrated HisPur Ni-NTA resin and incubated for 30 min at 4 °C with end-over-end mixing for batch binding. These were applied to gravity columns, allowed to drain, washed with 3 × 10 mL of wash buffer (lysis buffer with 40 mM imidazole), and eluted with 4 × 1 mL of elution buffer (lysis buffer with 250 mM imidazole).

For LAMP screening, elutions were pooled, dialyzed into storage buffer (10 mM Tris, pH 7.4, 100 mM KCl, 1 mM DTT, 0.1 mM EDTA, 0.5% Tween-20, 0.5% Triton-X100, and 50% glycerol), and stored at −20 °C. For thermoresistance characterization and RCA assays, elutions (Bst LF, Klentaq, v5.9) were further purified and instead dialyzed into buffer A (20 mM Tris, 150 mM NaCl, 1 mM DTT, 1 mM EDTA, and 0.1% Tween-20), then diluted with 14 mL of buffer A1 (buffer A without Tween-20). The eluate was applied to a gravity column with 1 mL of type I heparin agarose resin (Sigma) preequilibrated with 10 mL of buffer A1 and then washed with 2 × 10 mL of buffer A1. Proteins were eluted on a 0.15 to 0.8 M NaCl

gradient, with polymerases typically eluting at 470−575 mM NaCl. The elutions were pooled, dialyzed into storage buffer, and stored at −20 °C. For nickel affinity chromatography, the protein purity was 50−90% (typically ~80%), as indicated by SDS-PAGE electrophoresis. For sequential nickel and heparin affinity chromatography, proteins were ≥99% pure. For all assays, protein concentrations were equilibrated to commercial Bst LF (New England Biolabs) using SDS-PAGE densitometry, which has a concentration of 8 units/μL (1 unit is defined at the amount of enzyme that will incorporate 10 nmol of dNTP into acid-insoluble material in 30 min at 65 °C). We chose to normalize to concentration rather than activity in order to accurately compare the functionalities of our variants to those of their wild-type ancestor polymerases, Bst LF and Klentaq.

3.4.2 Real-time LAMP screening

LAMP reactions contained 1× Thermopol buffer (20 mM Tris-HCl, 10 mM (NH4)2SO4, 10 mM KCl, 2 mM MgSO4, 0.1% Triton X-100, pH 8.8) (New England Biolabs), an additional 2 mM MgSO4 (4 mM final concentration), 0.4 mM dNTPs, 20 pg of template (GAPDH, Table S2), 1× primer mix (FIP = 1.6 μM, BIP = 1.6 μM, LR = 0.8 μM, F3 = 0.4 μM, B3 = 0.4 μM; Table S2), 1 M betaine, 1× EvaGreen fluorescent DNA intercalating dye (Biotium) to monitor amplification, and 2.5 μL of polymerase in a total reaction volume of 25 μL. The reactions used five primers instead of the typical four or six as in previous studies[101]. Fully assembled reactions were heat-denatured at times and temperatures indicated in the text prior to LAMP with or without polymerase included (as noted) on a gradient thermal cycler. For SD Pol LAMP tests, SD Polymerase Hotstart (Bioron) was purchased, and reaction mixtures were assembled according to the manufacturer's recommendations (1× SD Reaction Buffer, 3.5 mM MgCl2, 15−50 units of SD Pol) with the addition of primers, template, and fluorescent dye as mentioned above; these reaction mixtures were heated for 2 min at 92 °C to activate hotstart before assaying.

The reactions were monitored using a LightCycler 96 quantitative PCR machine (Roche) by incubating at 68 °C and taking FAM fluorescence measurements every 4 min, followed by a post-amplification melt curve analysis to determine product specificity. The curves were analyzed using the accompanying software with absolute quantitation and Tm calling analyses. Thus, Cq values produced by the software (where indicated) represent crossing a

fluorescence threshold value determined by the software and correspond to a time point (multiply by 4 min) rather than a cycle number as in qPCR. Where present, error bars represent the standard error of the mean from three experimental replicates.

3.4.3 Thermoresistance assay

Thermoresistance assays were performed according to previously validated methods used to characterize Klentaq and Taq polymerases[90]. Kinetic activity assays were performed according to the manufacturer's instructions using the EvaEZ fluorometric polymerase activity assay kit (Biotium). The v5.9 polymerase (nickel- and heparinpurified) was diluted 1:4 from its working concentration and incubated at 85, 89.5, or 92.5 °C for 0, 1, 2, 5, or 10 min, followed immediately by snap cooling on ice. Polymerase (1 μL) was mixed with with 10 μL of 2× Polymerase Activity Mix and 9 μL of H2O and monitored on a Light Cycler 96 machine with readings every 30 s. Initial slopes, indicating reaction rates, were measured. Activities of samples were averaged across three experimental replicates for each temperature/incubation time, normalized to the no-heating control, and expressed as percentages. The error for each data point was derived from triplicate runs.

3.4.4 Rolling circle amplification assays

RCA reactions contained 1× Thermopol buffer , 0.4 mM dNTPs, 100 ng of pATetO plasmid template (excepting no template controls), 1× EvaGreen fluorescent DNA intercalating dye to monitor amplification, and 2.5 μL of polymerase (nickel- and heparinpurified) in a total reaction volume of 25 μL. Where indicated, reaction mixtures also included 20 primers (10 forward and 10 reverse, each 0.5 μM) for exponential amplification (JNM264–283; see Table S1). For reactions containing nicked template, 2.5 μg of plasmid pATetO was nicked in a 50 μL digestion reaction mixture containing 1× NEBuffer 3.1 (100 mM NaCl, 50 mM Tris-HCl, 10 mM MgCl2, 100 μg/mL BSA, pH 7.9) and 20 units of Nb.BsmI (New England Biolabs) by incubation at 65 °C for 1.5 h followed by 80 °C for 20 min to heat-kill the nickase. Aliquots (2 μL) of this reaction mixture or 50 ng/μL non-nicked pAtetO in 1× NEBuffer 3.1 were added to reaction mixtures containing template in order to maintain consistency.

Reactions were monitored using a LightCycler 96 machine by incubating at 68 °C and taking FAM fluorescence measurements every 4 min. Fluorescence measurements were exported from the instrument software and normalized to start at a fluorescence value of 0 at t = 0 by subtracting the initial fluorescence value from all subsequent measurements of a given reaction. The data set was then transformed by a multiplication factor of 1000 for easier visualization of the data. Maximum slope values represent the highest slope observed in a given curve between two measurements (4 min apart). All calculations were executed in Excel. Error bars represent the standard error of the mean from two experimental replicates.

3.4.5 Reverse transcriptase fidelity

Templates for single stranded consensus sequences (SSCS) were prepared by first strand reverse transcription or primer extension (plasmid DNA template) with a barcoded primer. Polymerization reactions were carried out according to manufacturer's recommendations for recombinant MMLV (New England Biolabs). For experimental polymerases, reverse transcription or primer extension was performed in 1x Assay Buffer, 200 μM dNTPs, 1 mM MgSO4, 400 nM barcoded reverse primer (HSP.seqBAR.R or pol2.SeqBar.R), 40 units RNasin Plus, 0.2 μg polymerase, and template (1 μg Human heart total RNA or 1ng plasmid). Reactions were incubated at 68◦C for 30 minutes (cDNA synthesis) or 2 minutes for DNA primer extension. Single stranded products were PCR amplified using Accuprime Pfx polymerase (ThermoFisher) with nextSeq.R and corresponding indexed forward primer. Samples were submitted for Illumina miseq PE 2x250. Targeted DNA sequencing reads were aligned and grouped based on unique molecular barcodes tagging individual reverse transcription events using ustacks (v1.35). Using a modified version of the SSCS program, only groups containing three or more reads were analyzed. From these reads, a consensus sequence was built if more than sixty-six percent of the bases at each position were in agreement, otherwise the base was called as N and disregarded in the remaining analysis. Consensus reads were then aligned to the reference sequence using BWA-MEM (v0.7.7), which allowed the detection of errors that derived from polymerase misincorporation as opposed to sequencing errors. Using this approach single nucleotide variants and indels were identified. The polymerase fidelity was

calculated as the sum of indels and erroneous bases as a fraction of the total number of aligned bases.

3.4.6 RT-PCR Assay

50 µL reverse transcription PCR (RTPCR) reactions were set up on ice with the following reaction conditions: 1x Assay Buffer, 1 mM MgSO4, 1 M Betaine (Sigma-Aldrich), 200 µM dNTPs, 400 nM reverse primer, 400 nM forward primer, 40 units RNasin Plus (Promega), 0.2 µg polymerase and 1 µg of Total RNA from Jurkat, Human Spleen or E. coli (Ambion). Primer sets used: PolR2A (PolII.R, PolII.F1/F2/F4), p532 (p532.R, p532.F1/F2/F5), rpoC (rpoC.R, rpoC.F1/F2/F4). Reactions were thermal-cycled according to the following parameters: 68◦C - 30 min, 25x (95◦C- 30 sec, 68◦C (63◦C for rpoC) - 30 sec, 68◦C - 30 s/kb).

3.4.7 RNA sequencing and analysis

RNA from U87MG glioblastoma cells (ATCC® HTB-14) were harvested using trizol LS following manufacturer's instructions (10296-028, Thermo fisher scientific). Ribosomal RNAs were then removed from the RNA samples using Ribozero rRNA removal kit (MRZH11124, Epicentre) and cleaned using RNeasy MinElute Cleanup Kit (Qiagen). rRNA depleted RNAs were fragmented using NEBNext Magnesium RNA Fragmentation Module (E6150S, NEB) to 200-300bp size range followed by kinase treatment to prepare for adaptor ligation. Illumina libraries were prepared using NEBNext Multiplex Small RNA Library Prep kit (E7580, NEB) and size selected to remove adaptor dimers using Ampure XP beads. 6 Illumina libraries were prepared from the same pool of RNA using experimental reverse transcriptases and ProtoScript II Reverse Transcriptase from the library prep kit. RNASeq libraries were sequenced on Illumina HiSeq 2000, 2x100bp by the genome sequencing and analysis facility at the University of Texas at Austin.

The evaluation of RNA-seq quality control metrics was performed via RNA-SeQC (v1.1.8). For transcript abundance analysis, fpkm values were generated through the cufflinks/cuffnorm pipeline (v2.2.1) and transformed both by log2 and to fit the range [-3,3].

Sanger sequencing reactions were set up by preparing 1x Assay Buffer, 1 mM MgSO4, 10 pmol RT.Probe, 50 pmol SangerGATC Template, 0.4 ug RTXexo- , and 50 µM dNTPs. For the indicated terminator nucleotide, a 25:1 ratio of 3' dideoxy terminator to unmodified NTP

was used. Reactions were thermal cycled 6x (68○C - 20sec, 85○C - 5sec). Reactions were terminated by the addition of EDTA to a final concentration of 25 mM. The labeled primer was removed by heating sample at 75◦C for 5 minutes in 1x dye (47.5% formamide, 0.01% SDS) and 1 nmol of unlabeled SangerBlocker oligonucleotide.

3.4.8 Encoding of information into oligonucleotides

We first combined each set of documents into a tar.xz file and padded the tail end with zeroes such that the final filesize was a multiple of 16 bytes. We then used DNA Fountain[143] to generate 4000 oligonucleotides encoding the cover message. We confirmed that none of these 4000 nucleotides had a DNA Fountain seed modulo 2000, a fact that will be used below to distinguish the hidden oligonucleotides from the cover set upon sequencing. For the hidden message, we first generated 2,000,000 DNA Fountain oligonucleotides, and kept only 2000 out of the 8933 whose DNA fountain seed was modulo 2000. We then computationally tested each oligonucleotide set, cover and hidden, to see how many sequences we required to recover each message. For each set, the oligonucleotides were shuffled into a random order and fed into DNA Fountain until the message was recovered (DNA Fountain terminates upon successfully recovering the message). This was repeated 1000 times. We recorded the number of oligonucleotides DNA Fountain required from each permutation before the message was decoded.

3.4.9 Synthesis of DNA and Omethyl DNA for Cryptogenetic Storage

The encoded oligonucleotide pools were each randomly arrayed on a 12,472 feature chip using the Customarray rearrayer software to give a ~3 fold sequence coverage for the standard unencrypted DNA pool (4,000 unique oligonucleotides) and ~6 fold sequence coverage for the encrypted 2'-O-Methyl-DNA oligonucleotide pool (2,000 unique oligonucleotides). The unencrypted DNA oligonucleotides were synthesized on the Customarray B3 oligonucleotide array synthesizer following standard phosphoramidite chemistry protocols. For the synthesis of the encrypted, 2'-O-Methyl oligonucleotides, 5 grams of each of the 2'-O-methyl phosphoramidites (2'-Ome Bz A, Cat. # 27-1842; 2'-Ome Ac C, Cat. # 27-1823; 2'-Ome U, 27-1825; 2'-Ome iBu G, Cat. # 27-1846) were purchased from Thermo Scientific and resuspended in 100mL anhydrous acetonitrile and used for

oligonucleotide synthesis on the chip following standard DNA synthesis protocols. Following the completion of the synthesis, the oligonucleotide pools were cleaved and deprotected directly from the chip surface using aqueous ammonia at 65°C for 4 hours. The cleaved and deprotected oligonucleotide pools were resuspended in TE buffer and purified on a Micro Bio-spin column (Biorad) following the manufacture's protocol. The column purified oligonucleotide pools were then used for further analysis.

3.4.10 Preparation of DNA for NGS Sequencing

Synthesized oligonucleotides were pooled in a ratio of 1 part DNA to 10 parts O-methyl DNA prior to amplification. To prepare oligonucleotides for NGS the pools were PCR amplified to add adaptor sequences. Reactions were indexed using Illumina small RNA primers (RPI1-KOD, RPI2-RTX, RPI3-RTX-Ome, RPI4-OneTaq One Step RTPCR (NEB)). For KOD, RTX, and RTX-Ome: 50µL PCR reactions were prepared with 1x Assay buffer, 200 µM dNTPS, 1 M Betaine, 400 nM RP1 primer, 400 nM RPI (1-3), 10 ng oligonucleotide pool, and 0.2 µg of KOD, RTX, or RTX-Ome polymerase (polymerase added after temperature reached 94°C). Reactions were cycled on using a program: 94°C - 30s; 12x cycles (94°C - 15s, 65°C (-1°C/cycle) - 15s, 68°C - 10 minutes). For OneTaq One-step RTPCR kit, the manufacturers recommended protocol was used with the same concentration of pooled oligonucleotides. After thermalcycling, products were cleaned using Wizard SV PCR purification kit (Promega) and eluted in 15 µL H2O. A secondary PCR was used to further amplify products from the RT-PCR before submission to the UT GSAF facility. Accuprime PFX PCR (Thermo Scientific) was used to amplify 5 µL of the eluted primary amplification with universal outnested primers (Universal F/Universal R) for 25 additional cycles.

3.4.11 Informatic recovery

Starting with raw sequencing reads, we first trimmed adapters and filtered reads to be between 50bp and 90bp using flexbar. We clustered the resulting reads using cd-hit at a 70% sequence similarity. For each cluster, we performed multiple sequence alignment using mafft with a gap penalty of zero and weighted bases according to the read's original length. A consensus sequence is built based off the most common base and gaps are filled until the sequence reaches our target length. Using knowledge of the Reed-Solomon code, GC content,

and homopolymer constraints, we ensured that the constructed consensus sequence matched the initial design parameters and if not, iterated through the gaps until such a sequence was found. Sequences were inputted into a modified DNA fountain program, where sequences needing less than 20 iterations were fixed and the remaining shuffled. The aggressive flag in DNA fountain was utilized and run 1000 times, with the most commonly occurring md5 checksum used as the basis for decoding.

3.5 Conclusion

This chapter first describes our work in creating a DNA polymerase chimera combining two unique properties—thermostability derived from Taq and strand displacing properties inherently found in Bst. Random recombination via gene shuffling was used to generate 10^7 chimeras with an average of 1.8 crossover events per variant. While this molecular biology technique has long been established, its use in conjunction with high throughput selections particularly in the polymerase realm is underdeveloped. Two novel variants were identified in rounds 5 and round 7 that could perform LAMP assays; however, only the variant from round 5 retained polymerase activity after a thermal challenge. Interestingly, a new capability, performing hbRCA from supercoiled plasmids, was found in this polymerase variant which neither parent enzymes could perform. Upon sequencing of this new variant, we found that the sequence was majority originating from Taq. The mutations responsible for endowing strand displacing activity arose primarily through a small 14 amino acid insertion extending a loop on a single helix in the thumb domain highlighting minor domains can confer unique properties.

The power of the method used in this chapter to derive novel properties comes from the agnostic approach to understanding functional residues. Rather than having prior knowledge of the strand displacing domain in Bst and subsequently try to rationally introduce the sequence into Taq, we instead relied on randomly combining the two genes and pulling out variants with the intended function using a novel selection scheme. As the unknown very much still prevails through biology, these mechanism-agnostic methods hold weight in delivering novel function where the mechanism of function is not known.

We next used a modified CSR method, but to inform the plasticity of a polymerase towards multiple evolution goals and show that an error-correcting thermophilic polymerase, KOD, is evolved to utilize two other nucleic acid substrates—RNA and 2'Ome. The culmination of both works highlight the extraordinary plasticity of the enzyme. RTX was created with only 16 single step mutations, amalgamated after paring down the mutations observed after next generation sequencing. Utilization for 2'Ome utilization was further evolved with another 11 mutations, two of which were reversions to wild-type KOD residues. In an enzyme with function critical to cellular viability, one could envision reaching an evolutionary cul-de-sac. However, the results presented here show that a degree of plasticity exists with perhaps a greater expansion of functionality than previously believed. Utility of RTX has been highlighted in this chapter and has since been used for immune repertoire analysis[158], RNA cycling, and low resource research settings[159]. RTX-Ome can provide a solution for information security in the burgeoning field of DNA storage.

Taken together, these studies underscore the notion that new functions within enzymes can be a relatively short evolutionary distance away. Owing to high-throughput screening methodologies, the feasibility of sampling large sequences can be accomplished in the lab in a manner that cannot be done in nature. In other chapters, we will see how laboratory evolution enforces organisms to adopt new translational machinery (Chapter 4) and how point mutations and short evolutionary distances can be more efficiently sampled using data driven methods (Chapter 2).

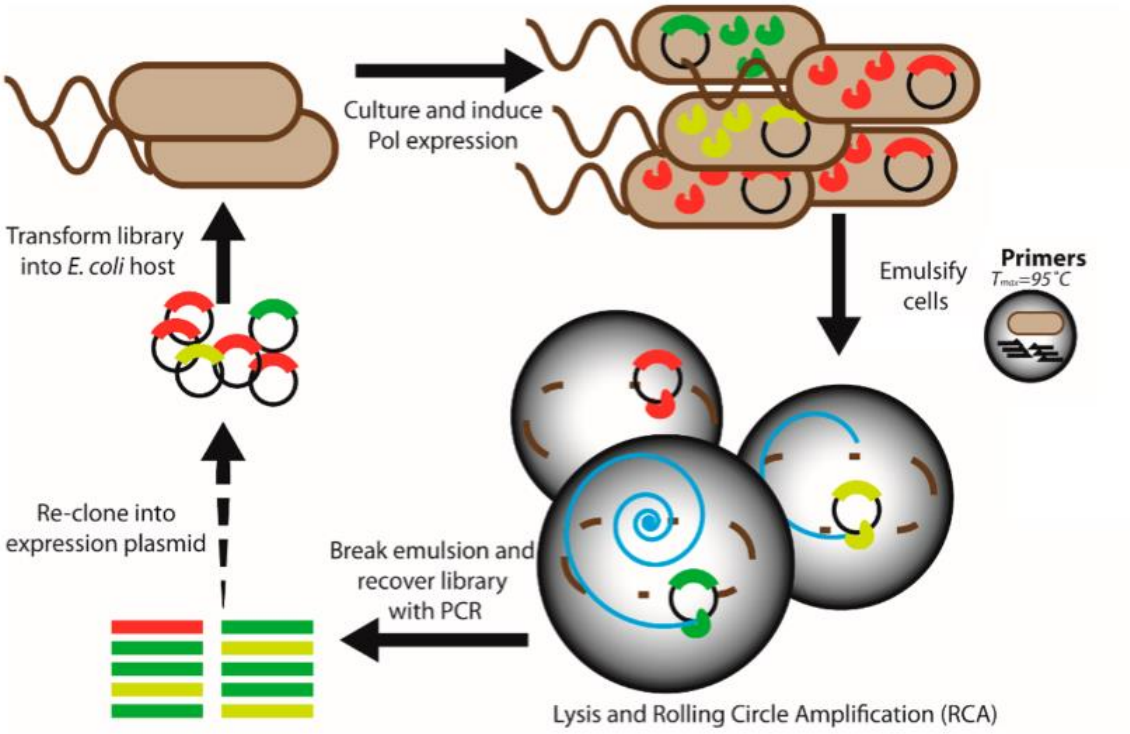


Figure 3.1 Schematic of HTI-CSR.

In HTI-CSR, E. coli cells expressing a plasmid-encoded polymerase library are suspended in a water-in-oil emulsion with a single cell per compartment, preserving the genotype–phenotype linkage. Each compartment contains primers for heat-initiated lysis and RCA in thermostable HTI-CSR. Following lysis, functional polymerases replicate their own plasmids via isothermal RCA at 65 °C, which is dependent upon a polymerase having strong strand displacement activity. The most active polymerases (green) produce more DNA, while less active variants (yellow) produce less; nonfunctional variants (red) produce none. After the hbRCA reaction, emulsions are broken, DNA is pooled, and the library is recovered by PCR, enriched for functional variants by the positive feedback loop. The library can be further recloned into the expression vector for subsequent rounds as needed.

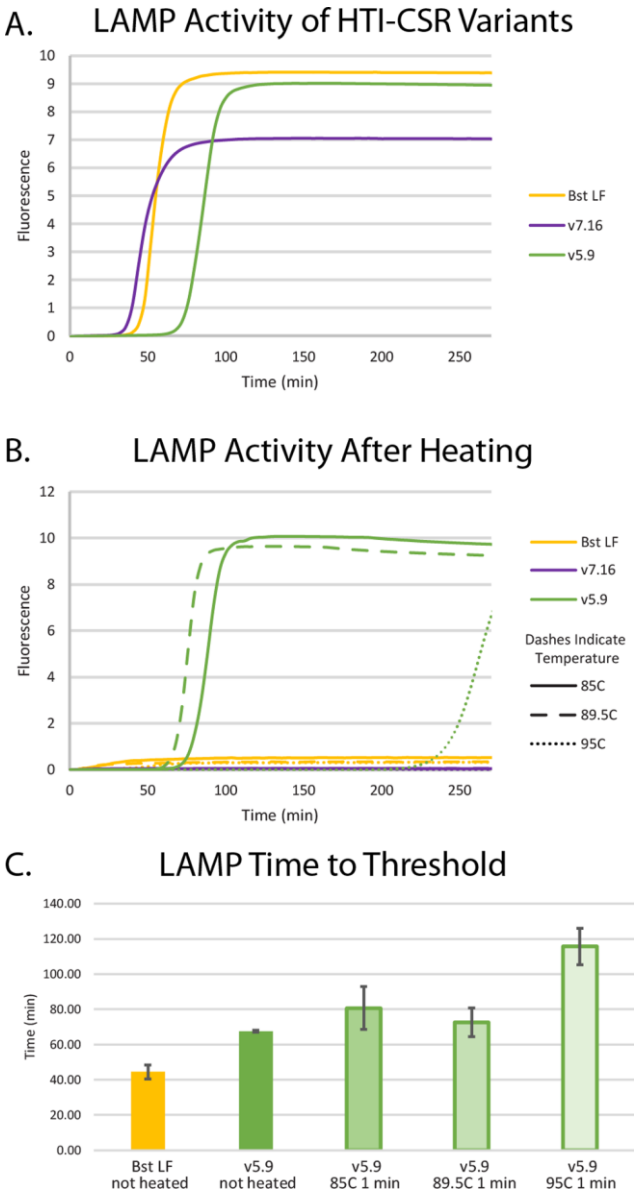


Figure 3.2 Isothermal screening of evolved variants.

(A) Real-time LAMP activities of two functional variants isolated from HTI-CSR compared with Bst LF. (B) Real-time LAMP activity after heating fully assembled LAMP reactions for 1 min at a range of temperatures. Solid and dashed lines indicate the temperatures at which the reaction was heated prior to LAMP. (C) Mean time to threshold for LAMP reactions with Bst LF or v5.9 with or without preheating of the enzyme. Each datum represents the average of replicates from three separate runs. Error bars represent the standard error of the mean. For v5.9 heated at 95 °C for 1 min, the average and standard deviation represent two replicates, as one replicate was not called positive by the software despite generating a curve (see B, v5.9–95C).

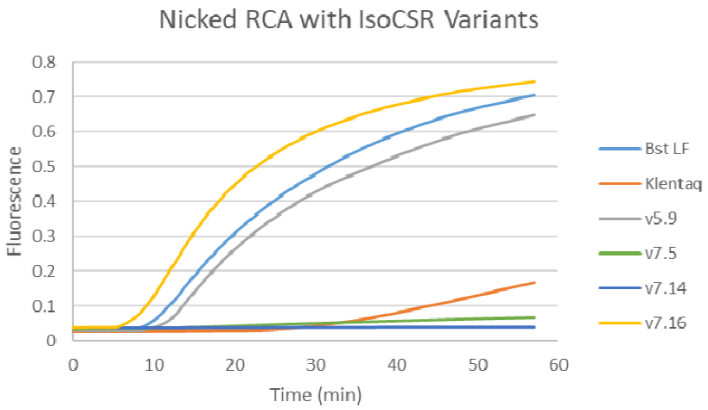


Figure 3.3 Nicked RCA reactions with variants isolated from HTI-CSR selections.

In addition to our stringent qLAMP screening, we also screened some variants with nicked RCA. These reactions included the nicking enzyme (Nb.BsmI, New England Biolabs) directly in the reactions rather than using pre-nicked plasmid template. While this type of screening still reveals highly functional polymerases like v5.9 and v7.16 seen in the main text, the increased resolution also identifies variants with limited function like v7.5.

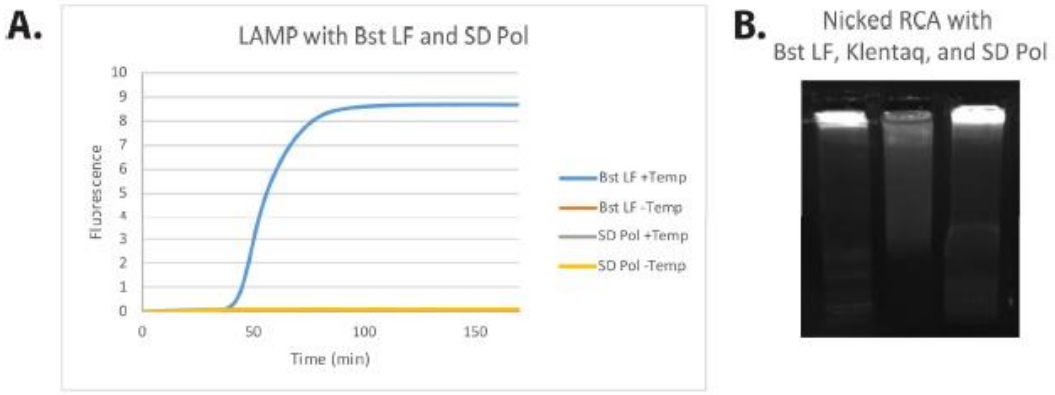


Figure 3.4 LAMP amplification with Bst LF and SD Pol.

We attempted to compare SD Pol (Bioron), a commercially available Taq mutant reportedly capable of LAMP (10), with our other variants in our initial qLAMP screening (A). Template is included as indicated. Using manufacturer recommended conditions, we were unable to generate amplicons with SD Pol. Other templates were also attempted (data not shown). SD pol is capable of Rolling Circle Amplification from a nicked plasmid template (B).

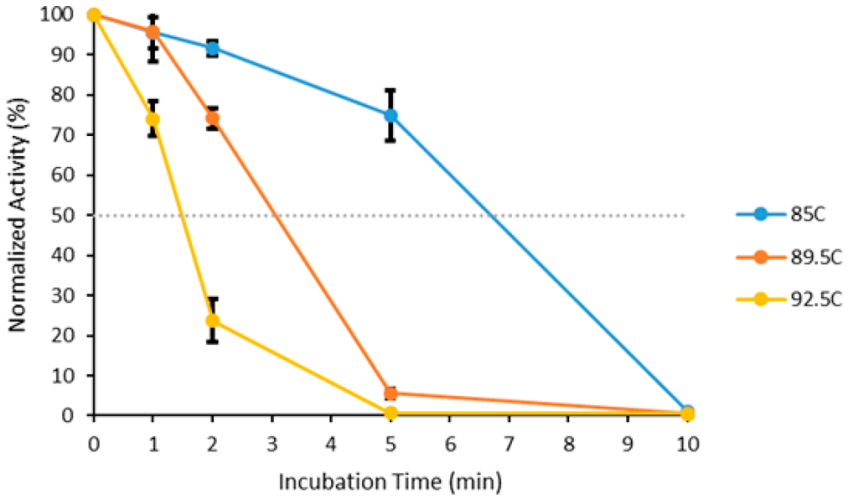


Figure 3.5 Thermoresistance kinetics of v5.9.

The thermoresistance of variant 5.9 was characterized by examining the polymerase's reaction rate in an extension assay after the polymerase was preheated at various temperatures and times. The activity for each time and temperature pair is averaged across triplicate runs and normalized to the enzyme activity without heating. Error bars indicate the standard error of the mean. A light-gray line indicates the half-life (50% activity).

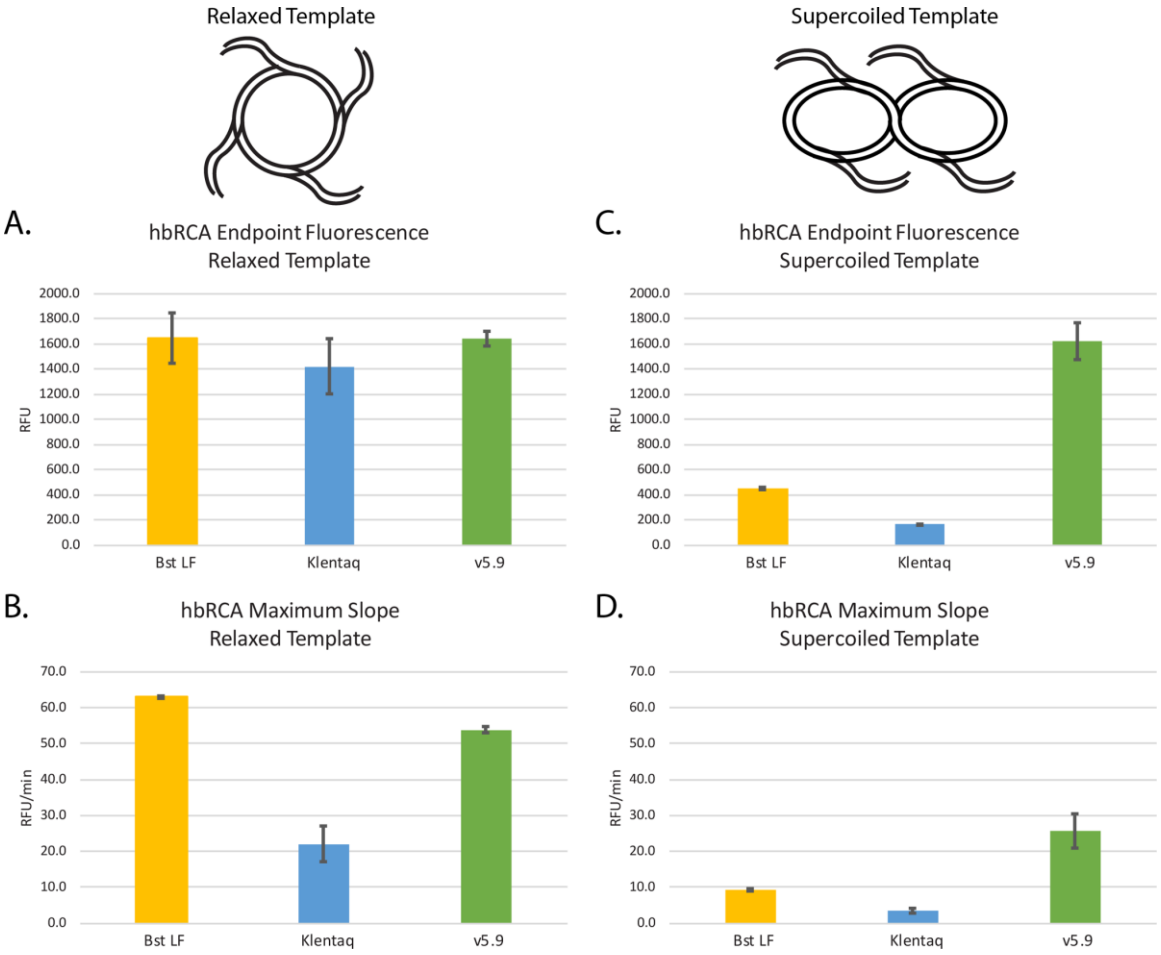


Figure 3.6 Hyperbranched RCA with v5.9, Bst LF, and Klentaq.

Enzymes were incubated with a nicked template with forward and reverse primers (A and B) or supercoiled plasmid with forward and reverse primers to mimic the RCA reaction in our selections (C and D). End-point fluorescence, indicative of reaction yield, is depicted in (A) and (C), while maximum slopes corresponding to reaction rate are depicted in (B) and (D). Polymerase colors are coordinated between graphs for comparison. Mean values across two experiments are indicated, with error bars corresponding to standard error of the mean.

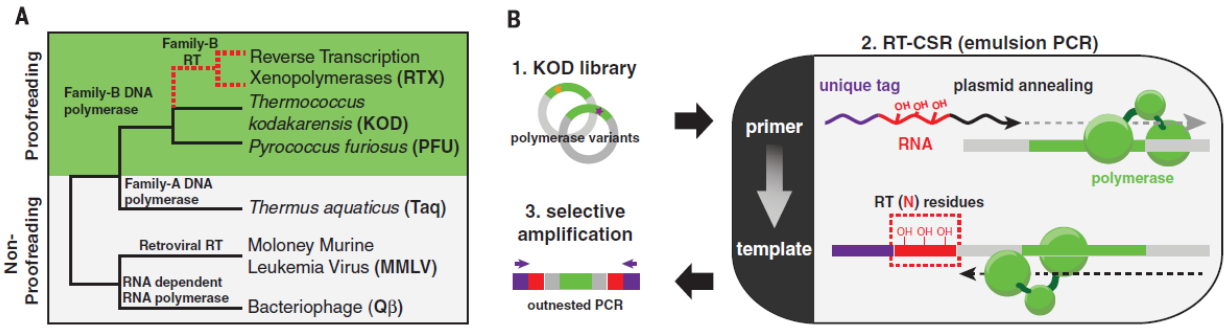


Figure 3.7 Evolution of a synthetic family of reverse transcriptases by RT-CSR.

(A) Polymerase phylogeny depicts reverse transcription xenopolymerases (RTX) as a second, evolutionarily distinct, origin of RT function. (B) Framework for the directed evolution of hyperthermostable RT using reverse transcription compartmentalized self-replication (RT-CSR). Libraries of polymerase variants are created, expressed in Escherichia coli, and in vitro compartmentalized. During emulsion PCR, primers flanking the polymerase enable self-replication but are designed with a variable number of RNA bases separating the plasmid annealing portion from the unique recovery tag.

Initial Selection			
Amino acid Position	Mutation Frequency	Amino Acid Change	Variant Frequency
97	94.2%	R -> H	58.70%
		R -> S	22.30%
		R -> C	13.20%
587	27.9%	F -> L	15.10%
		F -> L	12.80%
119		R -> H	10.90%

Round 10			
Amino acid Position	Mutation Frequency	Amino Acid Change	Variant Frequency
97	97.9%	R -> F	17.70%
		R -> A	11.80%
		Other	68.40%
384		Y -> H	81.20%
210		N -> D	63.70%
389		V -> I	50.20%
587	37.3%	F -> I	14.00%
		F -> L	23.30%
711		G -> S	29.30%
664		E -> K	29.20%
168		A -> T	25.70%
521		I -> L	24.20%
454		G -> D	22.20%
490		A -> T	17.40%
634		G -> D	16.00%
528		I -> L	14.50%
734		E -> K	14.10%
493		Y -> C	13.90%
311		Y -> C	12.10%
292		A -> T	11.80%
137		M -> I	11.30%
677		G -> S	10.90%
440		R -> H	10.80%
144		T -> A	10.80%
171		I -> V	10.60%
748		F -> Y	10.00%

Round 18			
Amino acid Position	Mutation Frequency	Amino Acid Change	Variant Frequency
384		Y -> H	96.00%
97	93.3%	R -> A	20.80%
		R -> F	18.00%
		Other	54.50%
389		V -> I	91.90%
210		N -> D	84.90%
493	83.3%	Y -> C	59.00%
		Y -> L	13.20%
		Y -> F	11.10%
664	82.7%	E -> K	60.40%
		E -> Q	22.30%
711	75.0%	G -> S	46.80%
		G -> V	28.20%
521		I -> L	59.40%
490		A -> T	58.50%
587	55.1%	F -> L	36.80%
		F -> I	18.30%
168		A -> T	36.70%
734		E -> K	34.50%
137	33.9%	M -> I	20.30%
		M -> L	13.60%
748		F -> Y	22.40%
735		N -> K	18.80%
593		K -> N	16.90%
590		T -> A	15.80%
605		T -> I	13.20%
143		E -> G	13.00%
501		R -> H	12.90%
144		T -> A	12.50%
150		E -> D	12.20%
145		L -> P	11.50%
741		V -> A	11.30%
692		K -> R	11.20%
454		G -> D	11.10%

Table 3.1 Deep sequencing of RT-CSR libraries.

Amino acid residues with mutations occurring in 10% of the population are shown in order of frequency. Some positions contained several amino acid possibilities and the sum of frequencies were totaled. Synonymous mutations are not shown.

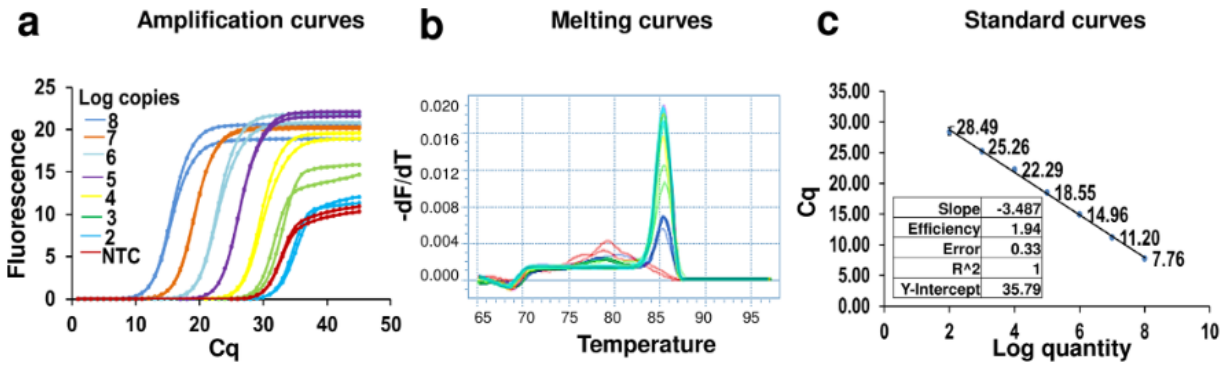


Figure 3.8 EvaGreen qRT-PCR analysis using RTX.

Indicated copies of synthetic Zika virus derived RNA template were amplified by RT-PCR using 80 ng of purified RTX polymerase. Amplicon accumulation was assessed in realtime by measuring increase in EvaGreen fluorescence. Representative amplification curves using 10^8, 10^7, 10^6, 10^5, 10^4, 10^3,10^2, 10^1, and 0 template RNA copies are shown in panel a. These curves were generated using the "Absquant" analysis protocol in the LightCycler 96 software. 'NTC' refers to no template control. The corresponding amplicon melting temperature analyses performed using the "Tm calling" protocol in the LightCycler 96 software are shown in panel b. The melting temperature of non-specific amplicons generated in the absence of templates is distinct from target-derived amplicons. Standard curve analyses performed using the "Absquant" protocol in the LightCycler 96 software are depicted in panel c. Standard curve analyses data for comparing amplification efficiency, linearity,and error are tabulated as insets.

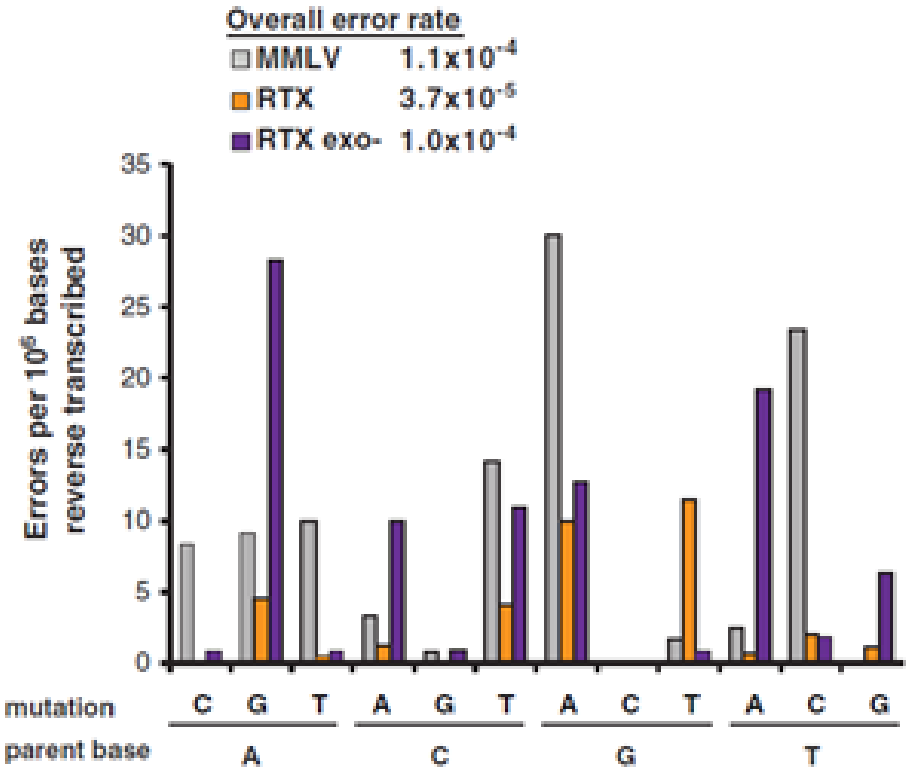


Figure 3.9 RTX polymerase proofreads during reverse transcription.

Deep sequencing of reverse transcription reaction on HSPCB gene. The overall error rate was determined by dividing the sum of base substitutions and insertions or deletions by the total number of bases sequenced.The error profile of MMLV, RTX, and RTX exo- is shown as frequency of errors per million bases sequenced.

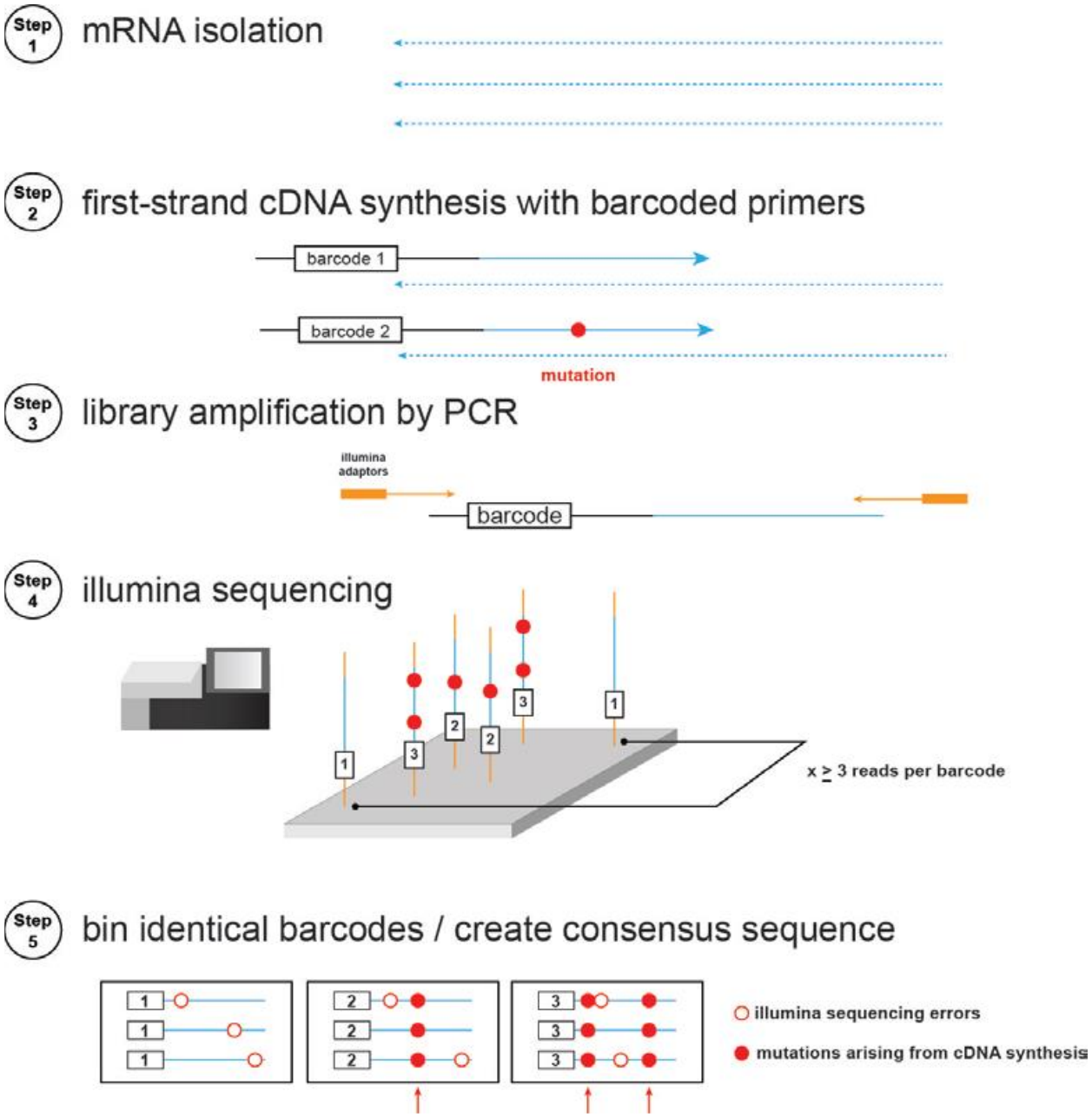


Figure 3.10 The SSCS method for reverse transcription.

In step 1, total mRNA is isolated. Step 2, barcoded gene specific primers are used to perform first strand cDNA synthesis. Step 3, the cDNAs are amplified with primers amplifying the cDNAs while preserving the barcodes. Step 4, Illumina miseq 2x250 paired end reads are performed enabling multiple reads of the same initial cDNA. Step 5, identical barcodes are binned and used to create a consensus sequence. Only barcodes that were read over 3 times were used in the alignment, reducing illumina derived mutations by >99%.

A

HSPCB reverse transcription

Polymerase	RTX	MMLV	RTX exo-	B11
Total Matches	1.44E+07	1.20E+06	1.10E+06	1.33E+07
Total Mismatch	520	124	102	3136
Total Indel	15	7	11	416
Error Rate	3.71E-05	1.10E-04	1.03E-04	2.66E-04
Base:Mutation	Mutation Frequency			
T to A	1.92%	2.42%	20.59%	23.82%
G to A	27.69%	29.03%	13.73%	6.12%
T to C	5.58%	22.58%	1.96%	2.68%
G to C	0.38%	0.00%	0.00%	0.00%
T to G	3.27%	0.00%	6.86%	8.16%
C to G	0.38%	0.81%	0.98%	2.36%
C to A	3.46%	3.23%	10.78%	6.92%
A to T	1.54%	9.68%	0.98%	1.66%
G to T	31.73%	1.61%	0.98%	4.11%
C to T	11.35%	13.71%	11.76%	10.01%
A to C	0.19%	8.06%	0.98%	0.70%
A to G	12.50%	8.87%	30.39%	33.45%

PolR2A reverse transcription

Polymerase	RTX	MMLV	RTX exo-
Total Matches	1.66E+07	1.12E+06	1.26E+07
Total Mismatch	537	536	4175
Total Indel	54	7	965
Error Rate	3.56E-05	4.86E-04	4.08E-04
Base:Mutation	Mutation Frequency		
T to A	2.61%	0.56%	35.52%
G to A	14.34%	1.68%	2.18%
T to C	13.04%	88.25%	2.68%
G to C	0.74%	0.37%	0.05%
T to G	1.49%	0.00%	1.51%
C to G	0.37%	0.19%	2.35%
C to A	6.89%	0.00%	5.27%
A to T	1.12%	0.19%	2.75%
G to T	34.08%	2.05%	3.83%
C to T	12.66%	2.80%	8.02%
A to C	0.56%	0.75%	1.20%
A to G	12.10%	3.17%	34.63%

B

HSPCB (DNA Template)

Polymerase	RTX	MMLV	RTX exo-	B11	KOD
Total Matches	1.84E+07	2.23E+06	4.65E+06	2.33E+07	1.49E+07
Total Mismatch	1521	297	795	5697	627
Total Indel	305	17	92	852	5
Error Rate	9.93E-05	1.41E-04	1.91E-04	2.80E-04	4.23E-05
Base:Mutation	Mutation Frequency				
T to A	4.67%	5.39%	15.47%	19.89%	2.71%
G to A	13.41%	14.14%	14.97%	9.60%	26.16%
T to C	3.35%	7.74%	3.40%	3.48%	5.74%
G to C	0.13%	0.34%	1.13%	2.42%	0.00%
T to G	0.66%	0.34%	2.14%	3.39%	0.00%
C to G	5.85%	1.35%	11.45%	11.01%	0.32%
C to A	14.73%	10.10%	16.48%	10.88%	34.13%
A to T	1.58%	12.46%	2.52%	6.90%	0.48%
G to T	6.38%	7.41%	2.64%	2.98%	6.54%
C to T	12.29%	8.08%	9.06%	10.67%	8.29%
A to C	0.72%	12.79%	0.75%	1.79%	0.80%
A to G	36.23%	19.87%	20.00%	16.99%	14.83%

Table 3.2 Fidelity for reverse transcription and replication.

(A) Fidelity profile for reverse transcription on two human genes, HSPCB and PolR2A. The error rate is calculated by dividing total mutations (mismatch + indel) by the total number of bases sequenced. The frequency of each possible mutation is listed as a percentage of total mutations. (B) Fidelity profile for DNA template (cloned plasmid DNA) polymerization using cloned HSPCB.

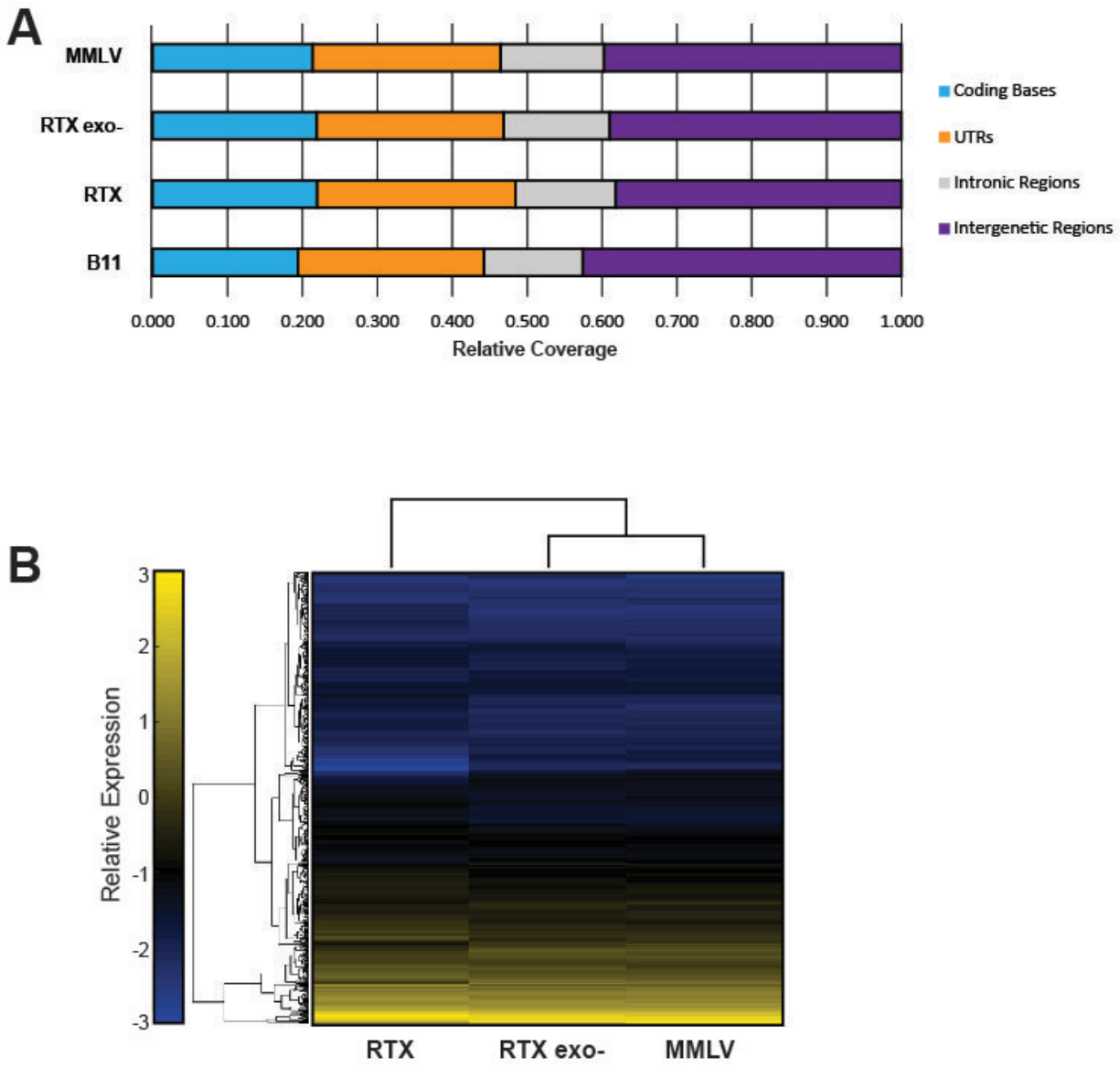


Figure 3.11 RTX in an RNA-seq workflow.

(A) Relative coverage of intracellular RNAs from gliobastoma cells for each reverse transcriptase. (B) Clustergram of relative expression for the top 500 most expressed RNAs for MMLV, RTX, and RTX exo-.

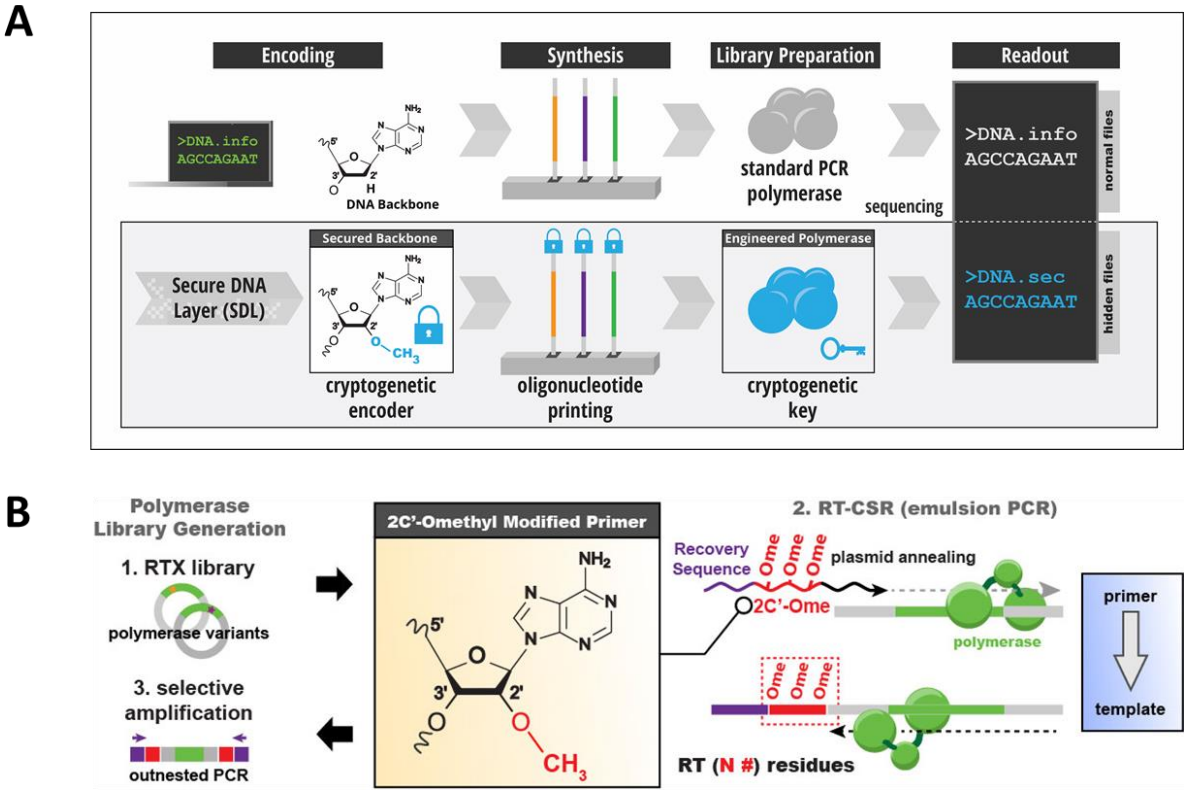


Figure 3.12 Evolution of a xDNA/polymerase pair creates a secire platform for DNA information.

A) Overview on cryptogenetics. Only with a paired DNA/xDNA pair can information be secured. B) Evolution strategy to create RTX-Ome

Amino Acid Position	RTX	Round 18	Variant Frequency
498	G	A	45.00%
251	E	K	41.80%
350	G	V	41.00%
159	M	T	33.60%
381	H	R	24.20%
488	I	L	22.20%
340	S	P	22.10%
384	H	Y	21.70%
468	K	N	21.60%
40	A	V	21.30%
353	V	L	20.20%
498	G	S	18.30%
289	K	R	18.20%
145	L	P	17.60%
242	Q	R	17.50%
664	K	R	17.20%
44	D	N	16.50%
244	M	F	16.10%
152	F	S	15.60%
418	V	I	15.30%

Amino Acid Position	RTX	Round 18	Variant Frequency
559	K	R	15.20%
276	E	D	15.20%
741	V	A	14.60%
484	R	H	14.30%
755	L	S	13.80%
168	A	T	13.70%
353	V	I	13.50%
768	W	R	12.30%
214	F	L	12.10%
247	R	L	11.50%
605	T	A	11.10%
704	L	I	10.90%
752	K	E	10.80%
640	V	I	10.80%
684	K	R	10.70%
703	V	I	10.60%
523	M	T	10.50%
248	F	L	10.50%
298	A	S	10.10%
309	A	T	10.00%

Table 3.3 NGS sequencing of the Ome RT-CSR Round 18 pool

Mutations are mapped to the parental RTX polymerase. Only mutations with over 10% frequency are shown.

A

Unsecured DNA		
File	Type	Size
AACS Encryption Flag	Image	1 KB
Budget	Excel	5 KB
Facebook Facial Recognition	HTML	2 KB
GPS Location History	KML	2 KB
Phone Contacts	Vcard	1 KB
Edgar Allen Poe "Gold Bug"	Text	30 KB
Tic Tac Toe	Game	1 KB
Tortilla Recipe	Text	1 KB

Secured DNA		
File	Type	Size
Cryptographie Indechiffrable	Text	3 KB
Schroder Message	Enigma Text	1 KB
Rasch Message	Enigma Text	1 KB
Kryptos Panel 1	Text	1 KB
Kryptos Panel 2	Text	1 KB
Kryptos Panel 3	Text	1 KB
Kyrptos Panel 4	Text	1 KB
Unbroken U Boat Message	Enigma Text	1 KB
Von Looks Message	Enigma Text	1 KB
Wikileaks Mission Statement	Text	12 KB
Zimmerman Telegram	Text	1 KB

B

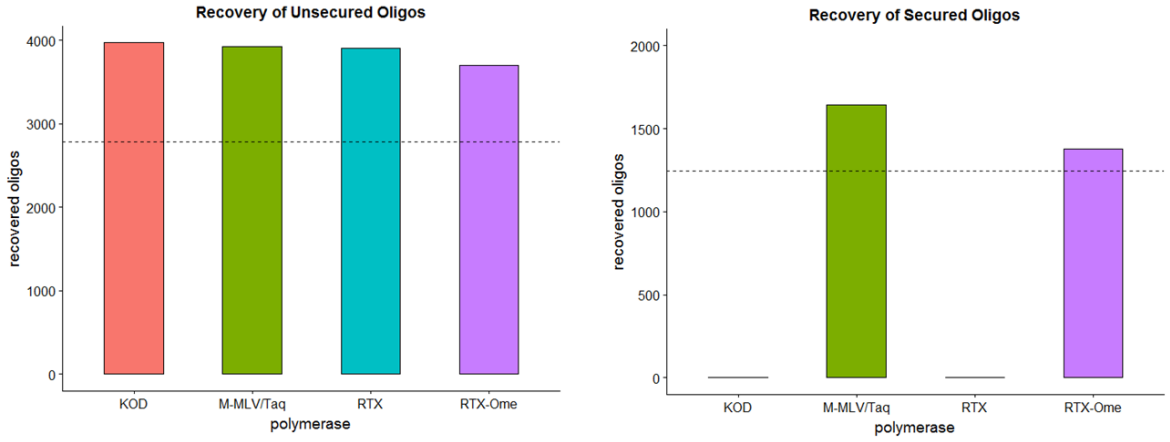


Figure 3.13 Encoding and decoding of information into oligonucleotides.

A) The listed files were encoded into DNA. B) Recovery performance of each tested polymerase in DNA (unsecured) and Ome (secured) oligos. The dotted line indicates the average number of oligos needed for decoding based on our simulations.

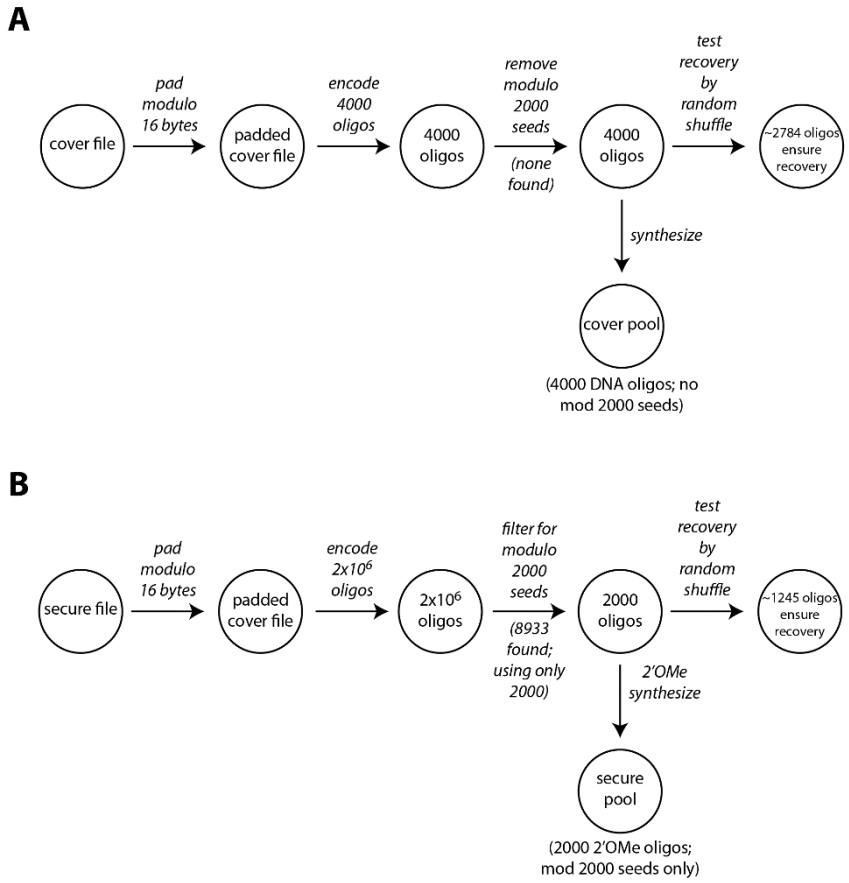


Figure 3.14 Information storage encoding schematic.

A) The cover file is encoded. First, it is padded to a multiple of 16 bytes for compatibility with DNA Fountain. We then let DNA Fountain generate 4000 oligos encoding it.

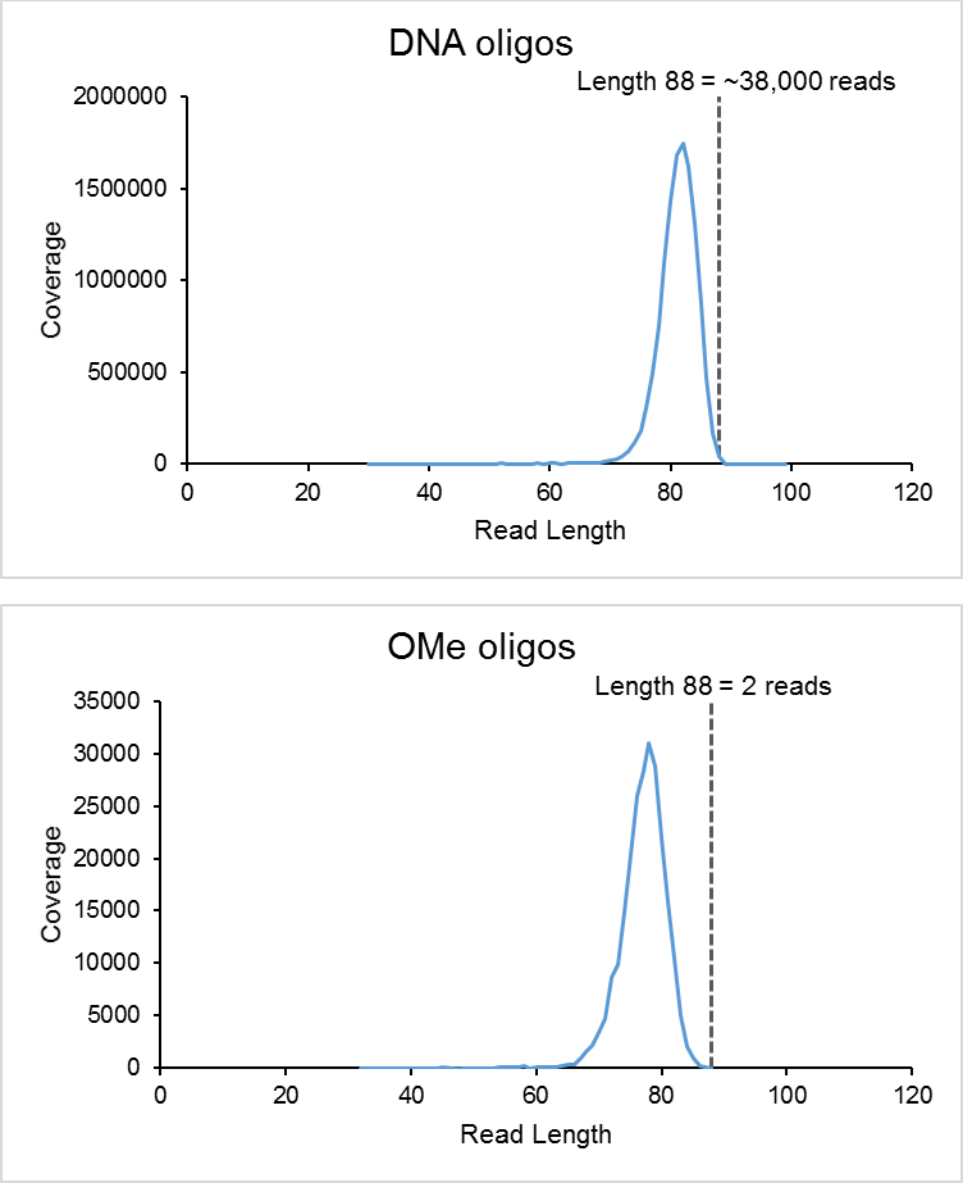


Figure 3.15 Distribution of NGS read sizes.

Following sequencing, analysis reveals the vast majority of sequences are less than the designed length of 88 bases.

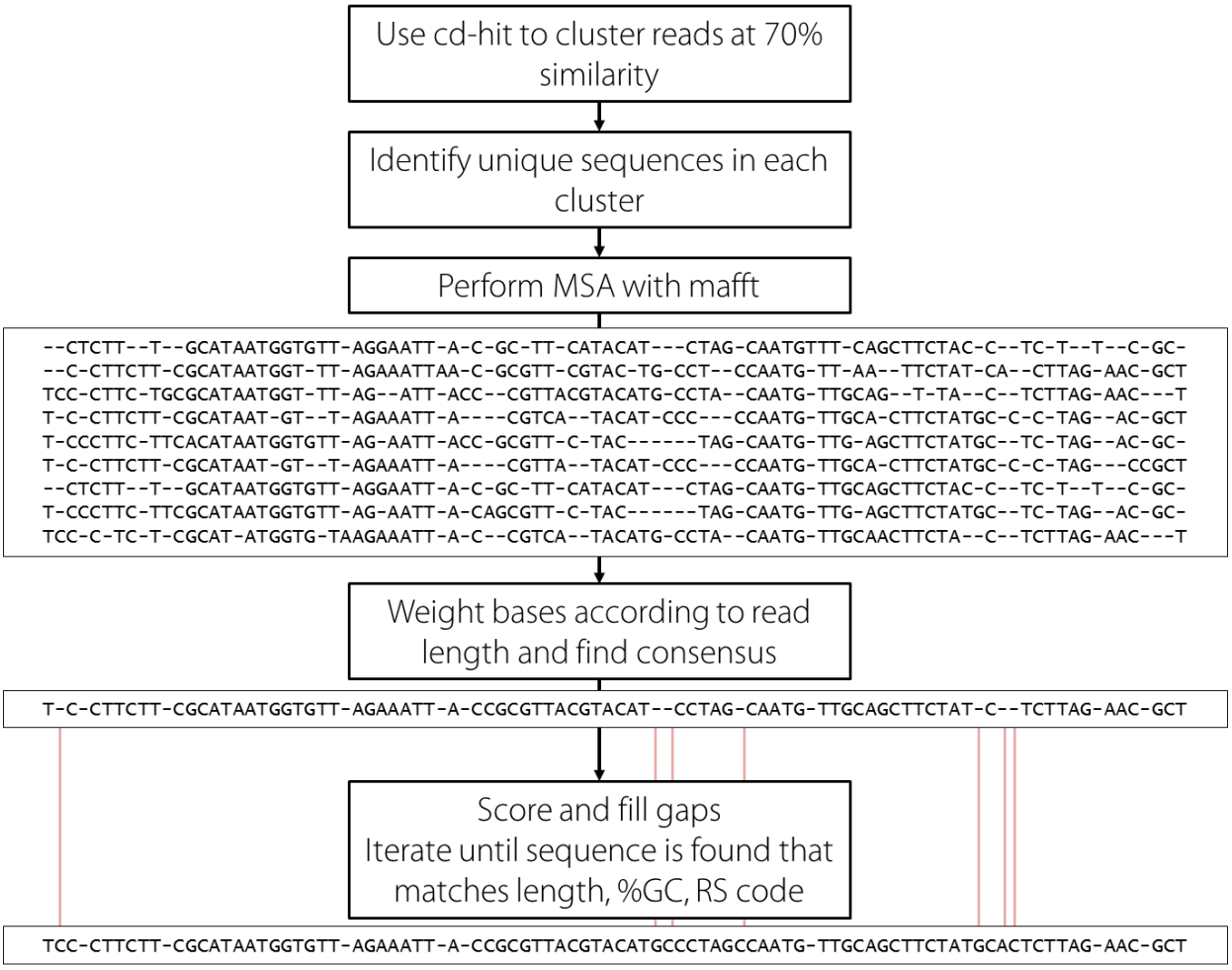


Figure 3.16 Decoding workflow of secured and unsecured oligos.

	Unsecure MD5 cd850ac885117191555adb3cbe34dd5b				Secure MD5 8253a5286479f5b87c0a63f073562380	
	KOD	RTX	RTX-Ome	MMLV/Taq	RTX-Ome	MMLV/Taq
Correct trials	564 *	1000	1000	1000	483 *	1000

***No other MD5 checksum appeared multiple times**

Table 3.4 Decoding trials.

Number of trials out of 1000, where decoding produced the correct md5 checksum for 4 tested polymerases. Of the incorrect checksums observed in unsecured KOD and secured RTX-Ome, none were observed more than once.

Chapter 4

Bacterial adaptations towards the adoption of expanded genetic codes

The genetic code has fixated to 20 amino acids. While this number seems somewhat arbitrary, the three branches of life (bacteria, archaea, and eukarya) largely obey this constraint. Amino acids outside the canonical twenty have the potential to introduce novel chemical interactions within a protein and recent efforts have expanded the genetic code through engineering of the underlying translational machinery and suppression of the amber codon. However, these systems often experience fitness deficits as proteins terminated by an amber codon can be unnaturally extended or biosynthesis of the new amino acid adds unintended toxicity. Retention of the orthogonal translation machinery can be enforced through an 'addiction' mechanism whereby the function of an antibiotic resistance element depends on the incorporation of the noncanonical amino acid.

This chapter explores two examples of how a cell might adapt to an expanded genetic code, the first two reports on a longitudinal study on 21st amino acid adoption. Both reports incorporate a noncanonical amino acid through suppression of the amber codon albeit in different strains. The first section centers on selenocysteine evolved in the context of an 'amberless' *E. coli*. The second section focuses on nitrotyrosine incorporation in an 'amberful' *E. coli*. Both experiments were grown for 2000-2500 generations, and the resulting genomic changes needed to efficiently accommodate a 21st amino acid were analyzed.

My contributions in both projects were to analyze the genomic changes that more efficiently enabled incorporation of each noncanonical amino acid. For selenocsyteine

---

This chapter is adapted in part from Thyer, R., Shroff, R., Klein, D. R., d'Oelsnitz, S., Cotham, V. C., Byrom, M., ... & Ellington, A. D. (2018). *Nature biotechnology*, 36(7), 624. It is presented with modifications under the full permission of the original publishers.

This chapter is also adapted in part from Tack, D. S., Cole, A. C., Shroff, R., Morrow, B. R., & Ellington, A. D. (2018). *Scientific reports*, 8(1), 3288. It is presented with modifications under the full permission of the original publishers.

My contributions are outlined in the text.

adoption, this included going a step further and making the identified mutations into the wild-type strain and characterizing the fitness effects. In the resulting manuscripts, I contributed to data presentation.

4.1 Introduction

Since the fixation of the genetic code evolution has been confined to the 20 canonical amino acids, with some incursions by selenocysteine and pyrrolysine. Alternative codon tables (e.g. mitochondrial genomes) are likely evolved from the standard codon table and provide evidence that the canonical genetic code can evolve[160]. A number of theories for the evolution of codon assignment and re-assignment have been proposed[161–163], and directed evolution experiments have demonstrated the code is not as frozen as once believed[164–166]. However, a full accounting of how a cell might adapt to an expanded genetic code has yet to be presented.

Expanding the standard set of proteinogenic amino acids can be accomplished through changes to the underlying translational machinery. Orthogonal translation systems (OTSs) comprising aminoacyl-tRNA synthetase (aaRS)/suppressor tRNA pairs have been developed that do not significantly interact with the host translational machinery or interfere with already occupied portions of the genetic code[167–169]. Typically, these OTSs allow the incorporation of noncanonical amino acids (ncAAs) by suppressing the amber stop codon (UAG).

Unsurprisingly, cells containing an active OTS often exhibit fitness deficits[170], possibly because any protein terminated by an amber codon can be unnaturally extended. Efforts to knockout the protein responsible for termination at amber codons, release factor 1 (*prfA*), support this claim: some strains lacking *prfA* were found to be viable only when essential genes terminating with an amber stop codon were recoded to terminate with an alternative stop codon[171]. In order to avoid these fitness impacts upon adopting a new code that would otherwise result in its rejection, previous studies with expanded genetic codes have either relied on bacteriophage, where the fitness of the host organism is irrelevant[166,172], or have relied on strains that entirely lack amber codons[173], allowing ready capture of the eliminated codon to create a 21 amino acid genetic code[174].

Here, we present two reports analyzing longitudinal adoption of expanded genetic codes. First, we explore selenocysteine which has long been exploited by nature for its unique biophysical properties, such as increased nucleophilicity and low pK_a and is a promising candidate for incorporation into engineered proteins by synthetic biologists. Previously, the Ellington lab evolved an *E. coli* $tRNA^{Sec}$ variant[175] that enabled efficient, site-specific selenocysteine incorporation at amber stop codons, but despite efforts to improve selenoprotein yield, encountered substantial toxicity from selenocysteine biosynthesis and low viability. Improving the recoded *E. coli* host (RTΔA[175], which is derived from C321.ΔA[176]) to the effects of genomic recoding[176] on regulation and synthesis of the proteome, could attentuate fitness burdens. Second, we explore the ncAA 3-nitro-L-tyrosine (3nY), which experiences similar fitness costs. This system models the 'ambiguous intermediate' hypothesis of genetic code evolution, which proposes that translation of a specific codon can change by first becoming ambiguously translated before losing ambiguity and gaining specificity for a different amino acid.

To improve fitness defects, we perform whole genome evolution, which has been used to optimize bacterial fitness, most notably in the laboratory evolution of *E. coli* over 60,000 generations by the Lenksi group[177] and resulted in complex genetic adaptations that supported the ability to metabolize a carbon source it couldn't previously use[178]. To facilitate the adaptation using genome evolution, we coupled cell survival to ncAA incorporation, by creating conditional dependence on an expanded genetic code and ensuring cells always exceed a minimum threshold of incorporation required for survival.

4.2 Results

4.2.1 Establishing selenocysteine dependence in E. coli

We produced a selenocysteine-dependent host strain by transforming *E. coli* RTΔA cells with two plasmids encoding a synthetic selenocysteine biosynthesis pathway consisting of $tRNA^{SecUX}$, selenocysteine synthase (SelA), selenophosphate synthase (SelD), and *O*-phosphoseryl-$tRNA^{Sec}$ kinase (PstK)[175]. Strains with different degrees of selenocysteine dependence were established by integrating one of three different variants of the $bla_{NMC\text{-}A}$ gene (disulfide-dependent β-lactamase from *Enterobacter cloacae*) into the chromosome[179].

The three variants were a wild-type β-lactamase containing the native disulfide bond between C69 and C238 (CC), which did not enforce dependence on selenocysteine, a U69-C238 variant (UC) containing an essential selenyl-sulfhydryl bond, and a U69-U238 variant (UU) containing an essential diselenide bond (**Figure 4.1a**). Selenocysteine-dependent β-lactamase activity of the UC and UU variants was confirmed by sensitivity to carbenicillin when grown in selenium-free defined media, which could be rescued by the addition of 1 μM sodium selenite (Na_2SeO_3). A wild-type *bla*$_{NMC-A}$ gene was also integrated into RTΔA cells containing empty plasmids (Δ strain) to provide a control to identify mutations that accrue independently of selenocysteine incorporation.

We selected for selenocysteine tolerance and improved fitness by serial passage of the four parental strains (CC, UC, UU, and Δ) for 2,500 cell doublings. All experiments were carried out in triplicate, under two different growth conditions that we hypothesized would elicit different adaptive responses. Growth of parental strains was either in an increasing concentration of a β-lactam antibiotic (β populations) or increasing temperature (T populations) (**Figure 4.1b**). These conditions were chosen because: (1) they are mildly mutagenic, through induction of DNA polymerase IV, thereby increasing the chance of an adaptive response[180,181]; (2) both conditions have previously been used, thus enabling us to compare our results with those already published; (3) both conditions impose stress on selenocysteine biosynthesis and incorporation, but in different ways. Increasing β-lactam antibiotic concentration exerts selective pressure for increased production of NMC-A β-lactamase, which in turn requires an increase in selenocysteine incorporation. Elevated temperature reduces $tRNA^{Sec}$ stability, which in turn decreases selenocysteine incorporation[182], and might result in evolution of more stable or more active selenocysteine incorporation machinery.

4.2.2 Genetic analysis of selenocysteine evolved strains

In order to analyze the genetic basis of evolved growth phenotypes, we carried out whole genome sequencing on evolved populations and their parental strains, including RTΔA. Single clonal isolates were also sequenced from each of the β_UC and β_UU populations and were found to be highly representative of the evolved population sequences,

containing the same fixed single-nucleotide polymorphisms (SNPs) and no genomic rearrangements. Whole genome sequencing indicated that the selenocysteine biosynthesis and incorporation machinery was maintained throughout evolution, that there was no loss of TAG codons in the *bla*$_{\text{NMC-A}}$ gene and no contamination between the populations. We did not find any genomic suppressor tRNAs via whole genome sequencing. We detected several new in-frame TAG codons in all sequenced genomes, including β_Δ and T_Δ populations, which contained no suppressor tRNAs (**Table 4.1**), likely decoded as glutamine as it is the most efficient near-cognate UAG suppressor[183].

Genome sequencing identified clusters of SNPs in genes conferring antibiotic resistance (*bla*$_{\text{NMC-A}}$, *ftsI*, and *ompR*), genes mediating oxidative stress or selenite resistance (*oxyR* and *cysK*), genes mutated during construction of C321.ΔA (*ftsA*, *hemA*, *pta*, and *yeeJ*), genes involved in plasmid replication (*pcnB*), and the *prfB* gene, which codes for release factor 2 (Figure 5.8a). Genes enriched with SNPs were defined as having acquired SNPs with >50% frequency in four or more independent populations evolved under the same conditions. Notably, no mutations were observed in the *selA*, *pstK*, tRNA$^{\text{SecUX}}$, or *selD* genes which form the synthetic selenocysteine biosynthesis pathway.

The impact of these mutations on fitness was evaluated by introducing a subset of the most enriched SNPs in genes not related to multidrug resistance back into the parental RTΔA strain, using multiplex automated genome engineering (MAGE)[184]. SNPs that we engineered into the parental strain included *oxyR* A233T, *cysK* T69I, T73A and H153Y, *prfB* T246A, N276D, and K282R, and also reversion of the mutations in *ftsA*, *hemA*, *pta*, and *yeeJ* to the wild-type sequence present in *E. coli* strain MG1655, the parental strain of C321.ΔA. The growth of MAGE-engineered strains was compared with that of RTΔA, and four mutations were identified that simultaneously decreased time spent in lag phase and increased the final culture density (**Figure 4.2b-e**). These mutations included three reversions of SNPs that were acquired during construction of the C321.ΔA strain, and the *prfB* T246A mutation, which repairs a defect in release factor 2 that is only present in *E. coli* K12 strains[185,186]. Although most of the SNPs we identified either compensated for metabolic defects in the parental recoded *E. coli* or broadly improved cell growth, *cysK* and *pcnB* directly affected selenocysteine tolerance and incorporation.

The number of mutations acquired by the populations (mean ± SD), defined as SNPs occurring in coding regions at >50% frequency, was highly variable (between 35 and 351), and while the T populations acquired more mutations on average (136 ± 83.8 versus 96.1 ± 35.6) the difference was not statistically significant. If the T_UC1 and T_UC3 populations which acquired mutations in *mutM* and seem to exhibit a hypermutator phenotype (acquiring 253 and 351 SNPs respectively) are excluded, the average number of mutations in the thermal populations decreases to 102 ± 26.1. The CC, UC and UU populations did not acquire more mutations than the Δ populations (97.4 ± 32.6 versus 103 ± 26.2), indicating that under these conditions selenocysteine biosynthesis was not inherently mutagenic.

Genome sequencing of the evolved populations revealed SNPs clustered in genes conferring antibiotic resistance, genes mediating ROS or selenite resistance, genes mutated during construction of C321.ΔA, genes involved in plasmid replication and the *prfB* gene encoding release factor 2. Genes which acquired mutations in more than four populations (and were present in >50% of the population) are listed in **Table 4.2** and **Table 4.3**.

Genes which may confer antibiotic resistance include *ompR*, which acquired the known activating mutation Y102C in 19 of the 24 evolved populations and regulates expression of the major porins (OmpC and OmpF) through which β-lactam antibiotics gain entry[187], and *ftsI* which encodes the penicillin binding protein III, the direct target of carbenicillin. Both the P311S and V545I mutations which were detected in *ftsI* in multiple populations have previously been reported to confer β-lactam resistance in *S. enterica*[188]. The integrated $bla_{\text{NMC-A}}$ gene was also a frequent site of mutation, with numerous SNPs observed at residues F105 and the N216-T217-T218 region. Mutations in these genes were not characterized further as the SNPs were observed in all populations as a specific response to the growth conditions, and have no bearing on the selenocysteine incorporation trait.

SNPs expected to affect ROS and selenite resistance included constitutively activating mutations in *oxyR* (A233T and likely C199R)[189] and *cysK* mutations predicted to abolish substrate and cofactor binding (P68L, T69I, T73A, R100H, F144S and H153Y) preventing the production of free selenocysteine[190]. To characterize the effect of these genes on cells fitness, the A233T mutation in OxyR and the T69I, T73A and H153Y mutations in CysK were introduced into the genome of RTΔA cells using MAGE and growth curves performed in the presence and absence of Na_2SeO_3 (**Figure 4.3**). All four mutants were either slightly

deleterious or no different from wild-type RTΔA cells under the various growth conditions. To confirm the growth assays, which were performed in 96-well plates, were representative of the culture conditions during the serial passaging experiment, we performed a competition experiment between the three *cysK* mutants and wild-type RTΔA cells. Cell populations containing 50% wild-type cells and 50% mutant cells were passaged in triplicate in LB media ± 10 μM Na_2SeO_3 for 125 generations and the change in frequency of the mutant alleles measured by qPCR (**Figure 4.4**). In agreement with the growth curves, the T69I mutant which was deleterious to cell growth in both the presence and absence of Na_2SeO_3 was lost from all cell populations (**Figure 4.4b**). Similarly, the T73A and H153Y mutants which were only deleterious in the absence of Na_2SeO_3, were only lost in these populations, and were retained in populations passaged in LB with 10 μM Na_2SeO_3 (**Figure 4.4d, f**). Despite strong enrichment of mutations in *cysK* over the course of the evolution experiment, most of which are known (or predicted) to abolish activity, these data do not provide a clear explanation of the selective pressures on the *cysK* gene or its role in selenite resistance.

In the case of OxyR, given mutations were only observed in populations exposed to β-lactam stress, it is possible that the transcription factor is mediating increased β-lactam resistance by enhancing expression of the $bla_{\text{NMC-A}}$ gene, rather than an oxidative stress response. While mutations in *oxyR* were also observed by Wannier et al. during continuous evolution of genomically recoded E. coli strains, the authors report that all mutations were functional knockouts[191]. This contrasts with the constitutively activating mutations observed in this work, indicating that they were derived under very different selection pressures. The $bla_{\text{NMC-A}}$ gene was integrated at the *ahpC* locus which is positively regulated by OxyR and much of the regulatory architecture upstream of the integration site is intact. Populations evolved under thermal stress also enriched for mutations in the *cpxR* and *iscR* genes. The *cpxR* gene encodes the regulatory protein of the CpxA-CpxR two-component system which modulates a variety of stress responses and also regulates expression of the OmpF porin which plays a role in β-lactam sensitivity[192]. The *iscR* gene encodes a repressor for the *iscRSUA* operon, of which *iscS* encodes cysteine desulfurase[193]. Cysteine desulfurase is a multifunctional protein responsible for selenide delivery during selenocysteine biosynthesis and also degradation of free selenocysteine.

Multiple SNPs were observed in four genes which acquired mutations during the MAGE process used to construct strain C321.ΔA[176]; *hemA*, *pta*, *yeeJ*, and *ftsA*. Of particular interest, three populations directly reverted the L196P mutation in *hemA* and four populations reverted the L673P mutation in *pta*, with a fifth acquiring a P673S mutation. To determine their influence on cell fitness we individually reverted the SNPs in the genome of the parental RTΔA strain (**Figure 4.5a-d**) to the wild-type sequence of MG1655. While no differences were observed in rich media for any of the genes, *hemA* and *yeeJ* significantly improved growth of RTΔA under more stringent growth conditions in defined media, and *ftsA* showed minor improvement. Interestingly, despite extremely high rates of mutation in *pta*, which was the most frequently mutated gene in both the β and T populations, we did not observe growth improvement when reverting P673 to the wild-type sequence. Deletion of *pta* is known to impair use of amino acids as a carbon source[194] and also survival in stationary phase, both of which are required during serial passage in LB media[195].

In recent work by Kuznetsov et al., out of the 355 SNPs known to have occurred during construction of the parental genomically recoded E*. coli* strain C321.ΔA, the authors reverted 127 'high-priority' SNPs back to the wild-type sequence and selected for combinations which improved cell growth rate[196]. In contrast, we found 49 highly enriched genes with sequence substitutions. Between these two sets of data, only *hemA* and *ftsA* overlap. Kuznetsov et al. found the *hemA* reversion to be one of the strongest drivers of improved fitness, and likewise it was highly enriched in our experiment. However, while neither Kuznetsov et al. nor we observed any growth improvement due to the *ftsA* reversion in rich media, we measured a mild fitness improvement under more stringent growth conditions in a defined medium. Thus, it is interesting that we also observed the accumulation of ten different second site mutations in the *ftsA* gene across nine independent populations. Similarly, in Wannier et al. the authors identify 52 genes which are enriched for SNPs during passaging on glucose minimal media[191], of which seven were also observed in our evolution experiment; *prfB*, *oxyR*, *pta*, *yeeJ*, *yneO*, *yfhM* and *rpoC*. Interestingly, the *pta* and *yeeJ* loci which were not reverted in Kuznetsov et al., were identified in Wannier et al., and were also found to be highly enriched for reversions and potential second site suppressors in our experiment. Taken together, the fact that different approaches identify different subsets of important

sequence substitutions with different functionalities highlights the complementary nature of our evolutionary approach to evolve functional protein-producing strains.

Evolved β populations also acquired mutations in the *polA* and *pcnB* genes which encode DNA polymerase I and poly(A) polymerase I respectively. Mutations in both genes are known to affect plasmid maintenance and copy number, including the S446F mutation which we observed[197]. Interestingly, the R105C mutation in the *pcnB* gene occurred independently during construction of the CC, UC and UU parent strains. This mutation is not present in the parental RTΔA strain, did not occur during construction of the parental control (Δ) strain and was not acquired during evolution of any of the control populations, which only contained empty plasmids. The *polA* S446F, T666A and M768V mutations and the R105C mutation in *pcnB* were introduced into the genome of RTΔA cells and their influence on plasmid copy number was measured by qPCR (**Figure 4.4g, h**). All mutations dramatically lowered plasmid copy number, consistent with literature reports for S446F. That the R105C mutation in *pcnB* occurred independently during construction of the strains containing plasmids pRSF-SelA-PstK and pCDF-SelD-UX indicates an extremely strong selective pressure to lower the toxicity or metabolic burden associated with highly active selenocysteine biosynthesis and/or incorporation.

Evolved populations also strongly enriched for mutations in the *prfB* encoding release factor 2 (RF2), particularly a T246A substitution, which was observed in 10 of the 24 populations. T246 is unique to K12 E. coli strains[185], with either alanine or serine found at that position in all other *E. coli* strains and related bacteria, and strongly decreases termination in response to UAA stop codons[186]. To validate the impact on translation termination, we introduced all observed RF2 mutations (T246A, T246A-N276D and K282R) into the genome of RTΔA cells and performed growth assays (**Figure 4.5e-h**). In addition, the N276D mutation was added individually. We observed a large improvement in cell growth rate and culture density when cells were grown in defined media for the T246A and T246A-N276D mutants and minor improvement for the K282R mutant. The N276D mutation alone was mildly deleterious and presumably co-enriched in cells which had acquired T246A. The *prfB* T246A mutation was also observed by Wannier et al. during passaging of recoded *E. coli* strains in glucose minimal media[191]. In agreement with our growth assay data (**Figure 4.5e-h**), the authors also report improved growth of strains carrying the *prfB* T246A

mutation in minimal media. Collectively these data support previous work which identified a requirement for efficient termination when *E. coli* cells were cultured in defined media with poorer carbon sources[186], and indicates that recoded *E. coli* strains suffer global defects in translation termination due to deletion of RF1.

4.2.3 Experimental evolution of bacteria in the presence of 3nY

We wished to examine the long-term adaptation and evolution of *E. coli* addicted to a ncAA, 3nY. We assembled an OTS for the incorporation of 3nY comprised of a *Methanocaldococcus jannaschii* tyrosyl-aaRS variant that had previously been engineered to be specific for 3-iodo-L-tyrosine[169] but was also compatible with 3nY[198], and the corresponding *M. jannaschii* tyrosyl-tRNA in which the anticodon was complementary to the UAG amber stop codon. This OTS enabled 'addiction' via a β-lactamase variant ($bla_{\text{TEM-1.B9}}$) that had been previously selected to be dependent upon 3nY incorporation at amino acid position 162[198]. However, since $bla_{\text{TEM-1.B9}}$ with 3nY already conferred resistance to high levels of ampicillin we further engineered $bla_{\text{TEM-1.B9}}$ to use ceftazidime (CAZ) as a substrate[199,200]. The new β-lactamase, bla_{Addicted}, conferred moderate resistance to CAZ in a 3nY-dependent manner at concentrations commonly used in bacterial cultures, with a measured minimal inhibitory concentration (MIC) of ceftazidime of 3–10 $\mu g\,mL^{-1}$ (**Figure 4.6**). This lower MIC allowed us to both retain and challenge the 3nY incorporation by progressively increasing CAZ concentrations during culture. We also constructed a control plasmid (pCONTROL) by replacing the 3nY codon (UAG) at position 162 in bla_{Addicted} with a phenylalanine codon (UUU), generating bla_{Control}. Phenylalanine is the only canonical amino acid that produces a functional β-lactamase when replacing 3nY162 in $bla_{\text{TEM-1.B9}}$[198]. As expected, the control lines conferred CAZ resistance in a 3nY-independent manner (**Figure 4.6**).

As a chassis for evolution, we chose to use *E. coli* strain MG1655 because it is well-characterized, with a sequenced and annotated genome[201]. MG1655 is autotrophic for all 20 canonical amino acids allowing for robust growth in amino acid knockout media. MG1655 was transformed with pADDICTED or pCONTROL, and lines were passaged in three different mixtures of amino acids in MOPS-EZ Rich Defined Medium (RDM). The first mixture contained all 20 standard amino acids (RDM-20), the second mixture lacked tyrosine (RDM-

19), and the third mixture lacked seven amino acids; serine, leucine, tryptophan, glutamine, tyrosine, lysine, and glutamate (RDM-13) (**Figure 4.7**). These seven amino acids represent all amino acids encoded by codons accessible through single nucleotide mutations from the UAG stop codon; by limiting the charging of the tRNAs for these amino acids, it should prove more difficult for any single mutation in a codon to be readily suppressed by mutations to tRNA anticodons or by mis-pairing. The RDM-13 media condition also proved a more stringent challenge to growth and evolution. Each media condition was supplemented with 10 mM 3nY, matching the concentration of L-serine, the most abundant amino acid in RDM.

We initiated selections with six independent clones, three containing pADDICTED, and three containing pCONTROL. These lines were denoted as Addicted(i), Addicted(ii), Addicted(iii), and Control(i), Control(ii), Control(iii) (**Figure 4.7**). Each clone was passaged in the three different amino acid environments described above, and evolved lines are identified by the evolutionary media condition and the progenitor clone from which it was initiated (e.g. Addicted-20(i), Addicted-19(i)…Control-13(iii)). Growth cultures were passaged daily by inoculating 5 mL of RDM with 1 µL of overnight growth. This resulted in approximately 12.5 generations per daily passage. While passaging, CAZ concentration was increased at a rate of 1 µg mL^{-1} per 100 generations to a final concentration of 22 µg mL^{-1} to provide evolutionary pressure and ensure enforcement of 3nY dependence. At the conclusion of 160 passages, corresponding to approximately 2000 generations, we selected a single clone from each evolved line and sequenced the bacterial genomes of the single clone as well as the entire bacterial culture using the HiSeq 4000 Illumina platform. Selected cells were, in general, genotypically representative of the average bacterial population. We further characterized phenotypic parameters of the selected evolved cells, as well as the progenitor cells.

4.2.4 Genetic analysis of 3nY adaption

During the course of passaging, CAZ concentrations were increased to levels beyond the MIC of the progenitor cells (the initial MIC of lines was approximately 2–6 µg mL^{-1}, while the final CAZ concentration challenge was 22 µg mL^{-1}). Bacterial survival at increasingly higher CAZ concentrations indicated that CAZ resistance was evolving in the lines. All nine of the addicted cell lines acquired at least a single mutation in *bla*$_{\text{Addicted}}$ during evolution

(**Table 4.4**). Several of these mutations are known or expected to be stabilizing mutations, while others are known to expand substrate specificity of $bla_{\text{TEM-1}}$[75,82,202–205]. Other mutations are specific to $bla_{\text{TEM-1.B9}}$, which originally included a number of substitutions relative to the wild-type $bla_{\text{TEM-1}}$ that addicted it to 3nY. The substitution T139I in Addicted-20(iii) and Control-19(ii), is more similar to the original wild-type residue, leucine, but does not reduce the 3nY dependence of the enzyme. In contrast, only five control lines acquired mutations in bla_{Control}: Control-20(i), -20(ii), -20(iii), -19(i), and -19(iii). The higher rate of mutation for bla_{Addicted} relative to bla_{Control} ($p = 0.0054$) indicates that the enzyme dependent upon the ncAA was not initially as fit as its non-addicted counterpart, especially in the presence of increasing antibiotic concentrations.

Beyond mutations in the *bla* gene itself, increases in MIC may have also been the result of other plasmid or genomic mutations that altered antibiotic resistance through other mechanisms. For example, the plasmid from Addicted-13(iii) has a mutation in *repC*, which can affect copy number and in turn antibiotic resistance[206,207]. In addition, genomic mutations occurred in genes that are known to have a role in antibiotic tolerance. Four lines (Addicted-20(i), -20(iii), -19(iii), and Control-20(iii)) had mutations in *envZ* (**Table 4.5**), a histidine kinase that regulates *ompF* and *ompC* expression, which in turn alter membrane porosity and have been tied to β-lactam resistance[208,209]. In the same vein, Addicted-13(iii) has a mutation directly in *ompF*. Mutations to *cyaA* or *crp*, two related proteins directly or indirectly involved in transcriptional regulation, occurred in eight evolved lines (Addicted-20(ii), -19(ii), -19(iii), -13(i), -13(iii) and Control-19(i), -19(iii), -13(iii)), and inactivation of these genes has been shown to produce resistance to β-lactams[210,211]. The *opgG* gene was mutated in Addicted-20(iii), -19(i), and Control-20(ii), and is involved in conferring resistance to antibiotics[212]. Finally, lipopolysaccharide (LPS) expression has been tied ceftazidime treatments, especially at or near MIC levels, and several lines had mutations in LPS biosynthesis and maintenance genes. Notably, seven lines had IS insertions in the *waa* operon that encodes the core oligosaccharide of LPS, including *waaO* (Addicted-20(ii), -19(i), and Control-20(ii), -20(iii)), *waaQ* (Control-20(i)), *waaP* (Addicted-13(i)), and *waaB* (Control-19(ii)). Also, three lines had mutations in the LPS-related gene *galU*, (Control-19(i), -13(i), -13(iii)), and one line had a mutation in *lptD*, which is involved in the assembly of LPS at the surface of the cell (Control-13(ii))[213,214].

Full genome sequencing and analysis revealed several trends among evolution conditions and the genes affected during evolution. (**Table 4.5**). Strikingly, the most commonly mutated or deleted ORFs were amino acid transporters in the hydroxyl and aromatic amino acid permease (HAAAP) family. The most commonly affected HAAAP protein was *tyrP*, a tyrosine-specific permease[215,216] that was inactivated or deleted in 10 of the evolved lines, including all six evolved in RDM-13, and four of the six lines evolved in RDM-19, including all three control lines, as well as Addicted-19(iii). The most common mechanism (9 of 10 cultures) used to inactivate *tyrP* was an IS1 mediated deletion of a large genomic fragment that excised twelve to fourteen genes, including a portion or the entirety of *tyrP*. The second most commonly mutated HAAAP protein was *mtr*, a tryptophan-specific permease[217]. The *mtr* mutations were found exclusively in lines evolved in RDM-13, with five of the six lines evolved in RDM-13 having a mutated *mtr* gene, and the final line, Control-13(iii), having an intergenic mutation near the *mtr* ORF. The third most common HAAAP protein that was inactivated was *sdaC*, a serine specific transporter, which was inactivated in four of the six lines evolved in RDM-20; a mutation of unknown consequence appeared in a fifth line (A384V). Surprisingly, *sdaC* was unaffected in all RDM-19 and RDM-13 evolved lines.

We hypothesized that deactivation of HAAAPs may have played a role in reducing the fitness burden of 3nY in the media. We sequenced the genomes of the six RDM-13 lines after 125 generations of growth in the enriched RDM-13 to determine which mutations had occurred early in experimental evolution, and found that five of the six populations had mutations in *mtr* or the *mtr* promoter. These mutations were fixed in three of the five lines, and for the remaining two lines some 22.8% of the Control-13(i) population had a *mtr* mutation, while 44.2% of the Addicted-19(iii) had mutations in *mtr*; both mutations were fixed by generation 2000. Additionally, the *tyrP* region was excised from the genomes of Addicted-13(i) and Control-13(ii). These mutations were the only identifiable mutations in the population after 125 generations, and were likely responsible for the ability to grow in pure RDM-13 after this point. This data supports the hypothesis that aromatic amino acid transporter deactivation (specifically *mtr* and *tyrP*) reduced the fitness burden imposed by the noncanonical aromatic amino acid, 3nY.

All eighteen lines were initially capable of utilizing 3nY in their proteome, and after 2000 generations of evolution four clones had each acquired a single in-frame amber codon in protein coding sequences. In three of these four instances, the strains did not require 3nY for growth. The clone from Addicted-20(iii) had acquired a SNP resulting in Q314tag in *opgG*, an osmoregulated peptidoglycan biosynthesis protein[218,219]. The other three in-frame amber codons appeared in clones from the Control lines, including Control-20(ii) and Control-19(iii), which acquired a W18tag mutation in *sdaC* and a W357tag mutation in *tyrP*, respectively. These amber substitutions likely down-regulated the expression of these HAAAP genes, consistent with our findings that these loci were frequently deleted or otherwise compromised. Control-19(iii) had deactivated the OTS entirely, indicating W357tag of *tyrP* was a truncation, while in Control-20(ii) the OTS remained partially active.

As mentioned above, two lines showed an unexpected dependence on 3nY for growth. When we assayed the ceftazidime MIC for the clone chosen from the Control-13(ii) line there was minimal to no growth on plates without 3nY, but appropriate growth on 3nY supplemented plates. Later, multiple replicates consistently resulted in a no growth phenotype for line Control-13(ii) in the absence of 3nY during amber suppression assays. Sequencing of this genome revealed an in-frame UAG codon in *lptD*, a Q557tag substitution. *lptD* is an essential gene involved in LPS biosynthesis[218,220]. Comparison to the full culture sequencing revealed this mutation was not representative of the bacterial population in the evolving culture Control-13(ii). Despite being a minor component of the full population, this mutation demonstrates the additional mutational space available to the evolving bacteria. An amber stop codon in *lptD* in a wild-type *E. coli* would result in a nonviable phenotype, yet with the expanded genetic code this clone has survived.

While the clone from the Control-13(ii) line was found to have an amber stop codon in an essential gene that may explain the lack of growth in the absence of 3nY, we cannot point to a similar rationale for why Addicted-19(ii) shows reduced growth rates in the absence of 3nY. Genome sequencing of the Addicted-19(ii) clone revealed six SNPs had arisen during evolution, four of which caused single amino acid substitutions, a fifth that resulted in a silent mutation, and a sixth that occurred in an intergenic region. As with other strains that had adapted to 3nY toxicity, Addicted-19(ii) contained a 10 kb genomic deletion surrounding the tyrosine permease *tyrP*.

4.4 Discussion

Previously, two different approaches have been used to generate organisms with altered or expanded amino acid genetic codes; a bottom-up approach where components of the organism were engineered to function with a ncAA, or a top-down approach, where an organism was allowed to evolve in the presence of a ncAA[221]. Advances in genome editing and protein structural modeling have made the bottom-up approach feasible, with two recent reports of *E. coli* successfully engineered to depend on a ncAA for survival[222,223]. It should now be possible to make fully synthetic genomes[224] that are designed to have a chemical dependence on an ncAA throughout the genome using the 'amberless' *E. coli*[176]. Alternatively, top-down approaches have previously been used to generate organisms that preferentially function with a ncAA substituting for tryptophan[164,225]. Using mutagens and selective growth conditions, *Bacillus subtilis* became dependent on a normally toxic tryptophan analog, 4-fluorotryptophan, with only 106 genomic mutations required to change amino acid preference[225]. In contrast, an *E. coli* tryptophan auxotroph evolved in the presence of 4-fluorotryptophan in place of tryptophan[226] could tolerate high levels of the ncAA, but in the end still required tryptophan for growth. A similar approach using long-term evolution in *E. coli* allowed for cells to grow using a sulfur-containing tryptophan analog in place of tryptophan[165]. Another recent report shows that the bacteriophage T7 will adopt ncAAs into its genetic code to reach new fitness peaks[166].

The work described herein can be considered a hybrid approach, resting between bottom-up and top-down. We used a single engineered addiction element, to preserve an active OTS over evolutionary timeframes within an otherwise unmodified organism. We first engineered an improved bacterial host for selenoprotein expression by endowing *E. coli* strains with conditional dependence for incorporation of selenocysteine using an amber suppressor $tRNA^{Sec}$ variant and an engineered β-lactamase containing an essential diselenide bond. Dependence on selenocysteine incorporation was maintained for more than 2,500 generations, allowing cell populations to evolve tolerance to constitutive selenocysteine biosynthesis and acquire several mutations that improved cell fitness. This is the first demonstration, to our knowledge, of enforcing an expanded genetic code to facilitate host adaptation and improve recombinant protein production. We anticipate that selected

mutations identified in our study could be integrated into recoded *E. coli* strains[196] that have been optimized using different methods. With recent advances in orthogonal translation systems[227] and protein design, this type of approach should prove generalizable to other amino acids in other proteins[198,222,223].

Previous attempts to improve genomically recoded *E. coli* strains have focused exclusively on growth rate[191,196] but have found less substantial improvements in the suppression of UAG codons by non-canonical amino acids. Using our evolved strains, we have greatly expanded the number of selenocysteine residues that can be site-specifically incorporated into recombinant or endogenous proteins, enabling the production of selenoproteins containing multiple diselenide bonds, such as seleno-antibody fragments. Evolved β_CC and β_UC populations showed five- to sevenfold increases in selenoprotein expression compared to the parental cells. Evolved β_UU populations showed less change from the parental cells, which had much higher fluorescence than either the β_CC or β_UC parental cells. In addition to the increased fluorescence (normalized to cell density), the evolved β_UC and β_UU strains achieved several fold higher cell density , increasing overall selenoprotein yield in some populations up to 20-fold.

Diselenide bonds have been used to improve the stability and biological half-life of therapeutic peptides, such as insulin[228] and oxytocin[229], which are manufactured using solid-phase synthesis. The ability to efficiently introduce diselenide bonds into recombinant proteins extends this stabilizing motif to previously inaccessible classes of protein therapeutics, and provides a high-throughput approach to improve existing therapeutics.

In addition to enabling the incorporation of diselenide bonds as a protein engineering tool[230], a reliable host for recombinant selenoprotein expression will find broad utility among the wider protein biology community. Replacing catalytic cysteine residues in enzymes with selenocysteine has enabled advances in mechanistic enzymology[231,232], but progress has been hindered by inefficient protein expression in cysteine auxotrophs or by specialized protein ligation strategies[233]. In addition, these approaches have inherent limitations, requiring either the removal of native cysteine residues or accepting indiscriminate selenocysteine incorporation, or maintaining the solubility of truncated enzyme fragments.

Similarly, efforts to characterize the human selenoproteome (comprising 25 proteins of which half are uncharacterized) have relied on auxotrophic selenocysteine incorporation[234], or replacement of catalytic selenocysteine residues with cysteine to overcome the difficulty of expression at the cost of producing proteins with low activity and unknown biological relevance[235]. Recent attempts to produce two of these proteins have been successful using complete chemical synthesis of entire selenoproteins[236]. Expression in a bacterial host will enable easier analysis of these proteins. Furthermore, selenocysteine-dependent conjugation chemistries[237], methods currently employed by peptide chemists, can now be expanded to recombinant proteins using our strains, making orthogonal drug conjugation easier by removing the need to eliminate other reactive surface residues. We envision that the tools and host strains for highly efficient site-specific selenocysteine incorporation that we report here will serve as a platform for exploring the potential of the selenoproteome and seleno-stabilized therapeutics.

Over the described 2000 generations, 3nY-addicted bacteria have remained dependent on 3nY for survival in evolutionary conditions, and cells have evolved to overcome the fitness burdens initially seen from the expanded genetic code. This was not a foregone conclusion: in order to re-establish the canonical 20 amino acid code, lines could have evolved bla_{Addicted} to no longer require 3nY, or acquired genomic mutations leading the CAZ resistance without bla_{Addicted}. Even though 326 genomic amber codons were a single mutational step from stop codons that would not be suppressed by the OTS, in over 2000 generations addicted populations fixed no mutations that led away from amber stop codons used for translational termination, despite previous evidence that recoding amber codons to alternative stop codons reduces the fitness effects of obligate UAG suppression[166].

Instead it became clear that the most important adaptations to the ambiguous genetic code we had imposed were ones that alleviated either the toxicity of an unnatural amino acid and/or that better optimized the composition of the 20 amino acids normally used for growth. Growth in 3nY clearly led to fitness deficits, and to compensate if the media lacked tyrosine (RDM-19 and RDM-13), the tyrosine permease *tyrP* was inactivated, and if the media also lacked tryptophan (RDM-13), the tryptophan permease *mtr* was inactivated. In contrast, if all 20 amino acids were present (RDM-20), the serine permease *sdaC* was inactivated. Since serine is the most abundant amino acid in RDM conditions (10 mM), it is

possible that the amino acid pool is better balanced through deletion of *sdaC*, while still relying on other serine transporters (*sstT*) for serine uptake.

Once strains had evolved to the point where they could accommodate an ambiguous genetic code that could accept suppression with either tyrosine or 3nY, they were positioned for further evolution to specify a new code. After 2000 generations of evolution with 3nY, the existence of three in-frame amber codons in clones with active OTSs, and the fact that a clone from the Control-13(ii) line had adopted an in-frame UAG codon in the essential gene *lptD*, provides evidence that populations are exploring the 3nY mutational space. Additionally, the reduced doubling time of a second line, Addicted-19(ii), in the absence of 3nY may indicate some preference for the new media condition containing 3nY, even though no genomic in-frame amber codons had arisen. In this regard, the evolution of Addicted-19(ii) may resemble the evolution of a *B*. *subtilis* strain that could preferentially utilize 4-fluoro-tryptophan in place of tryptophan[164].

Overall, our method provides one of the first experiments investigating how a new genetic code is adopted by an organism, and evolved lineages may represent evolutionary intermediates to the adoption of a new amino acid.

4.5 Methods

4.5.1 Molecular biology

To perform MAGE, 100 µL of RTΔA cells containing pKD78 from a saturated culture was diluted into 3 mL of LB supplemented with 33 µg/mL chloramphenicol and grown to mid-log phase at 30 °C. To induce the lambda red-recombination machinery, 100 µL of 10% w/v L-arabinose was added to give a final concentration of 0.3% and the cells transferred to 37 °C and incubated for 1 h. One mL of the induced culture was removed and centrifuged at 8,000*g* for 1 min to pellet the cells. Cell pellets were resuspended in 10% glycerol and washed three times to prepare electrocompetent cells. Mutagenic oligonucleotides were added to a final concentration of 1 µM each. Cells were electroporated at 1.8 kV, 25 µF capacitance and 200 ohms (Bio-Rad *E. coli* Pulser) and recovered in three mL LB supplemented with chloramphenicol. Cells were grown to mid-log phase and an additional two MAGE cycles were performed as described above. Following a final 3-h recovery, tenfold

serial dilutions were plated on LB supplemented with 33 μg/mL zeocin to obtain single colonies. Mutants were identified using multiplex allele-specific colony (MASC)-PCR, after which the target genes were amplified by PCR and mutations confirmed by Sanger sequencing.

4.5.2 Bacterial passaging and growth assays

To passage bacterial cells, 5-mL cultures of LB supplemented with 25 μg/mL kanamycin, 50 μg/mL spectinomycin, 10 μM Na_2SeO_3, and 100 μg/mL carbenicillin were inoculated in triplicate with the Δ, CC, UC, and UU parent strains. Following growth to stationary phase, cells were diluted 5,000-fold into fresh media resulting in ~12.5 doublings every passage. Glycerol stocks were prepared from all cultures every five passages and the selection stringency was increased every ten passages. During the β-lactam resistance experiment, stringency was adjusted by increasing the carbenicillin concentration by 100 μg/mL. This continued evenly until 2,500 cell doublings had occurred. For the thermal tolerance experiment, the temperature was increased by 0.5 °C (and the carbenicillin concentration kept constant at 100 μg/mL). At an incubation temperature of 43.5 °C, some cultures could no longer be passaged reliably when the freshly diluted cells were immediately incubated at 43.5 °C. To overcome this problem, all cultures in the thermal experiment were pre-incubated at 43 °C for 3 h and then the incubation temperature elevated to the correct level. At 45.5 °C several cultures had poor viability even with a 43 °C pre-incubation, and all remaining passages were performed at 45 °C. For the competition experiment between wild-type RTΔA cells and those carrying mutant *cysK* alleles, strains were initially grown to saturation, diluted to OD_{600} 0.1 and mixed in a 1:1 ratio. 5-mL cultures of LB supplemented with 33 μg/mL zeocin and ± 10 μM Na_2SeO_3 were inoculated in triplicate with 1 μL of the RTΔA:mutant cell mix. Cultures were incubated at 37 °C and serially passaged to saturation ten times (125 generations).

Parental and evolved *E. coli* strains were characterized by growth assays in both rich and defined media. For characterization in rich media, 5-mL cultures of LB containing 25 μg/mL kanamycin, 50 μg/mL spectinomycin, and 10 μM Na_2SeO_3 were inoculated from glycerol stocks and grown to saturation. Aliquots of each culture were diluted to OD_{600} 1.0 and 1 μL used to inoculate 100-μL cultures, in triplicate, comprising four different

carbenicillin concentrations in a 96-well plate. The growth assay media consisted of LB containing 25 μg/mL kanamycin, 50 μg/mL spectinomycin, 10 μM Na_2SeO_3, and the four carbenicillin concentrations were 0, 100, 1,000, or 10,000 μg/mL. Plates were sealed with an optically clear, gas-permeable membrane and incubated with constant orbital agitation (amplitude of 3 mm) at 37 °C. OD_{600} measurements were taken at 5-min intervals for 24 h (Tecan Infinite M200 Pro). Populations evolved with β-lactam stress were assayed using all four carbenicillin concentrations. Populations evolved with thermal stress were only assayed at 100 μg/mL carbenicillin. All growth curves are plotted as the mean of three biological replicates performed in technical triplicate ± s.e.m. represented as ribbon.

For characterization in defined media, cultures were started in MOPS EZ containing 0.5 μg/mL D-biotin, 25 μg/mL kanamycin, and 50 μg/mL spectinomycin. Parental strains were observed to grow poorly in MOPS EZ and cultures were diluted to OD_{600} 0.1 rather than 1.0. Growth assays were performed as previously described using four variations of the MOPS EZ media used for overnight growth; media only, media containing 100 μg/mL carbenicillin, media containing 1 μM Na_2SeO_3 and media containing both carbenicillin and Na_2SeO_3. Mutant strains generated by MAGE were assayed as described above in either LB supplemented with 33 μg/mL zeocin or MOPS EZ containing 33 μg/mL zeocin and 0.5 μg/mL D-biotin. Na_2SeO_3 was supplemented at 10 μM in LB and 1 μM in MOPS EZ. Growth curve data are representative of two or three repeated experiments.

4.5.3 Whole genome sequencing and bioinformatic analysis.

Genomic DNA from RTΔA cells, the parental Δ, CC, UC, and UU stocks and each evolved bacterial population was extracted from ~5 × 10^9 cells using a Zymo Research Fungal/Bacterial DNA Kit according to the manufacturer's instructions. DNA was prepared for sequencing with a 300 base-pair target insert size using standard methods. Samples were sequenced on an Illumina HiSeq 2500 system using 125 base-paired-end reads at the Genome Sequencing and Analysis Facility (University of Texas at Austin). Across all sequenced bacterial lines, an average coverage of 144.4× was obtained with a s.d. of 43.0. Raw sequencing reads were processed by trimming and removing adapters using trimmomatic (v0.32). The sequence of the RTΔA genome was assembled using the *De novo* Assembly Module in Geneious. Variant detection was performed using breseq (v0.27.0), with

the assembled RTΔA genome and plasmid sequences as references. Mutations occurring at or above 20% in the Δ, CC, UC, and UU parental strains were removed in their respective evolved populations. Genes enriched with SNPs were defined as having acquired SNPs with >50% frequency in four or more independent populations evolved under the same conditions.

4.5.4 qPCR and mutant allele detection

RTΔA cells along with the *polA* and *pcnB* mutants were transformed with the pRSF and pCDF empty plasmids. Replicates of each of the three clones for each mutant/plasmid pair were grown to saturation and then diluted 1:100 in a 96-well plate. Plates were incubated in a plate reader (BioTek Cytation 5) at 37 °C for 4 h with constant agitation, where OD_{600} was monitored. Following incubation the plates were removed and put on ice, where a 2-µL aliquot from each well was added to a qPCR reaction mix containing EvaGreen double-stranded DNA dye, and qPCR primers specific for either the *aphA1* or *aadA* genes from pRSF and pCDF, respectively. Purified plasmid DNA was quantified, diluted, and used as a standard on each qPCR plate for quantification. Fluorescence was read and analyzed using a Roche Lightcycler 96. Absolute quantifications of each sample were normalized to the OD_{600} of the well corresponding to the sample.

For detection of mutant *cysK* alleles, cultures were inoculated from glycerol stocks taken from the final serial passage for each sample and incubated at 37 °C with 225 r.p.m. agitation. For samples passaged with selenite, the medium was supplemented with 10 µM Na_2SeO_3 for any further growth. Saturated cultures were diluted 1:50 in LB and incubated for 3 h until reaching mid-log phase. 1.5-mL aliquots of each culture were normalized to the highest OD_{600} (~0.4) and centrifuged at 3,500*g* for 5 min, then resuspended in 100 µL of LB. Cells were then boiled for 15 min at 95 °C to prevent PCR inhibition. Cell debris was pelleted by centrifugation and the supernatant was recovered. Primers specific for wild-type and mutant alleles at T69, T73, and H153 were designed as previously described[238]. Oligos were purchased from IDT (Coralville, IA). Triplicate qPCR reactions (20 µL) were set up using 500 nM Forward and Reverse primer, 10 µL of 2× Fast EvaGreen qPCR Master Mix (Biotium, Inc., Fremont, CA), and 2 µL of cell supernatant from each sample. Reactions were run on the

Roche LightCycler 96 and analyzed using the manufacturer's software. qPCR data shown are representative of three experiments.

4.5.5 Statistical analysis and reproducibility

All data in the manuscript are displayed as mean ± s.e.m. unless specifically indicated. Bacterial growth curves and bar graphs were plotted in R 3.1.2 using the package ggplot2. The ELISA curves were estimated in R using a general asymmetric five parameter logistic model with the package drm and plotted using ggplot2.

4.5.6 Evolutionary set up

Plasmid backbone from pMMB67EH[239] was amplified using DNA oligos (Integrated DNA Technologies) DT01 and DT02. The *M. jannaschii* OTS and $bla_{\text{TEM-1.b9}}$ were amplified from a plasmid described previously[198] using oligos DT03 and DT04. To convert the penicillinase $bla_{\text{TEM-1.B9}}$ to the cephalosporinase bla_{Addicted}, residues 165–167 were converted from WEP to YYG using oligo DT05 with either DT06 or DT07 for bla_{Addicted} or bla_{Control} respectively. Reaction mixtures were transformed into *E. coli* TOP10 and selected on LB-agar with 2 μg mL^{-1} CAZ for bla_{Control}, and the same conditions with 10 mM 3nY for bla_{Addicted}. Samples were sequence verified at University of Texas core facilities using Sanger sequencing. Properly sequenced plasmid of bla_{Control} and bla_{Addicted} were used as pCONTROL and pADDICTED respectively, and transformed into *E. coli* MG1655. Three colonies from each were selected as clones i, ii, and iii for passaging.

MOPS-EZ Rich Defined Media (RDM, TEKnova) with the full complement of amino acids (RDM-20), as well as the knockout medias (RDM-19 and RDM-13), were prepared according the manufacturers specification. For media preparation, 3-nitro-L-tyrosine or 3-iodo-L-tyrosine (Sigma-Aldrich) was added to ultrapure deionized water to a final concentration of 17.24 mM. The 3nY or 3iY supplemented water was used to complete RDM, and the entire preparation was filter sterilized with Nalgene Rapid-Flow SFCA filtration units. Prepared media was stored at 4 °C, and moved to room temperature 16–24 hours before use.

Selected colonies i, ii, and iii, from each transformation were picked and grown in RDM-20 supplemented with 10 mM 3nY and 2 μg mL^{-1} CAZ for 16 hours. Each culture was then

used to inoculate three subcultures of RDM-20, RDM-19, and RDM13. Initially, cultures were incapable of growth in RDM-13, but were capable of growth when 25% of media was replaced with RDM-19. This supplemented RDM-13 was used for the first 125 generations for RDM-13 evolved cultures, after which cultures were capable of growth in RDM-13. Cultures were passaged every 16–24 hours by transferring 1 µL of culture into 5 mL of fresh media, and grown shaking at 37 °C. A 500 µL sample from each line was preserved at generation 0, 125, 250, and every 250 generations for the duration of evolution, samples are stored in 25% glycerol at −80 °C.

4.5.7 Genome sequencing and assembly

After the 2000 generations, 1 µL from each line was streaked onto RDM-agar supplemented with 10 mM 3nY. Two single colonies from each line were selected and grown in 3 mL RDM-20 with 10 mM 3nY with 22 µg mL^{-1} CAZ. Simultaneously, samples from the progenitor cells were grown in similar conditions using 2 µg mL^{-1} CAZ. After overnight growth, glycerol stocks were made of clonal cultures, and genomic DNA was isolated from one of the two cultures using bacteria genome miniprep kit (Sigma-Aldrich). Genomic DNA from mixed cultures were also prepped. Genomes were sequenced using the HiSeq. 4000 platform, 150 bp paired ends, achieving greater than 100× coverage across all samples. Raw reads were processed through trimming and adapter removal using trimmomatic (v0.32)[240]. Alignment of sequencing reads and variant calling was performed through the breseq workflow (v0.27.2)[241].

4.5.8 Data availability

All data generated or analyzed during this study are available from the corresponding author on reasonable request. Sequence data is available as NCBI, BioProject ID PRJNA430697.

4.6 Conclusion

Here, we show two reports of improved bacterial fitness in response to the addition of new translation machinery. While the orthogonal translation system and bacterial context

differed between the two experiments presented in this chapter, we did observe notable similarities. Both long term evolution experiments enforced retention of the OTS through an ncAA dependent beta-lactamase, and mutations were observed within this gene showing selection pressure on the antibiotic. In nitrotyrosine, the mutations were demonstrated to improve stability, but the mutational role observed in selenocysteine evolution was unclear. Both long term evolution experiments experienced toxicity burdens with unique solutions. To ameliorate the metabolic cost tied to selenocysteine incorporation and synthesis, *E. coli* cells reduced plasmid copy number. Addition of nitrotyrosine to the growth media was toxic, and the cell's response was to inactivate the amino acid transporters. In addition, mutations in both studies were observed in the EnvZ/OmpR regulatory system, a modulator of osmolality and the global stress response.

Mutations in the translation machinery were also common among the two experiments, though the likely effects both increased and decreased functionality of the translation apparatus in order to balance the complex system. Incorporation of nitrotyrosine in non-addicted cell lines mutated the suppressor tRNA to alleviate fitness burden. The selenocysteine dependent lines improved release factor 2, presumably to accommodate the increased ochre and opal burden. While grown for ~2500 generations, both evolved lines did not readily adopt the new amino acid or take advantage of the new chemistry provided. Selenocysteine addicted *E. coli* showed few mutations to amber, though these likely arose from drift and are overall, neutral. The new ambers in the evolved nitrotyrosine addicted cells follow a similar pattern, save for one which inserted into the essential gene, *lptD*.

While nonstandard amino acids have been engineered, the associated burdens and resulting fixes had not been studied in depth nor longitudinally. This chapter provides answers in two parallel experiments. With a top-down approach, we rely on laboratory evolution for solutions, letting the nature find optimality to a challenging task. Unlike Chapters 2 and 3 where the solution space is confined to a single gene, here we sought changes at a genome-wide scale. Many of the mutations crucial to increased fitness can be rationalized after identification through next generation sequencing, but would be near impossible to do from the onset of the experiment.

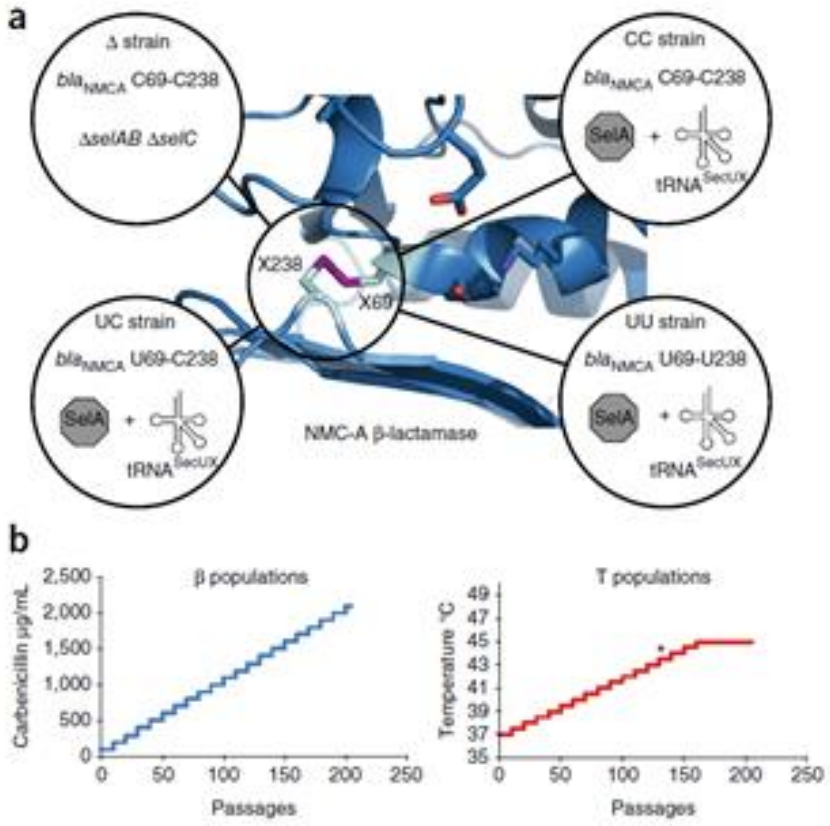


Figure 4.1 Evolution of selenocysteine-dependent E. coli strains improves fitness.

(a) A recoded E. coli strain deficient in selenocysteine incorporation (RTΔA) was made conditionally dependent on selenocysteine by integrating NMC-A β-lactamase variants containing either an essential selenyl-sulfhydryl (UC) or diselenide (UU) bond and supplying the biosynthesis and incorporation machinery in trans. Control strains contained a wild-type β-lactamase containing a disulfide bond, lacking either selenocysteine dependence (CC) or the entire incorporation trait (Δ). (b) The four populations were serially passaged, in triplicate, 200 times to saturation (~2,500 generations) in two different environmental conditions: increasing β-lactam stress (β populations) or increasing temperature (T populations). Note, at 43.5 °C (*) passaging technique was adjusted for the T populations. Curves are plotted with a line showing the mean and shading representing ± s.e.m., where n = 3 independent biological replicates.

gene	T_Δ1	T_Δ2	T_Δ3	T_CC1	T_CC2	T_CC3	T_UC1	T_UC2	T_UC3	T_UU1	T_UU2	T_UU3
xseB				Q50*								
mhpA				Q68*								
ygcB						W406*						
araD						Q8*						
yfeR						Q31*						
yhdP						W511*						
hfq					Q41*							
ftsK					**Q767***				**Q648***			
macB	Q20*											
mdtM	Q410*											
fetA		Q9*										
pgpB			Q152*									
yahG			Q85*									
rfbA			Q210*									
rapA			Q952*									
sbmA			W99*									
ilvB			Q68*									
wzc									Q553*			
hrpA									Q495*			
yfcU									W381*			
slyA									W16*			
bcsE									W482*			
opgB								Q261*				
ecpD								Q393*				
proA							Q295*					
marR											W83*	
yffS											W89*	
yfbK												Q301*
murF												**W224***
folA												**W30***
wecC												Q398*
purL												Q166*
mutM									W157*	Q212*		
mlrA										Q225*		
betT										Q580*		
yedF										Q45*		

Table 4.1 Amber mutations in evolved lines.

Genes containing in-frame TAG codons enriched in several populations. Gene marked in blue are essential for *E. coli* viability.

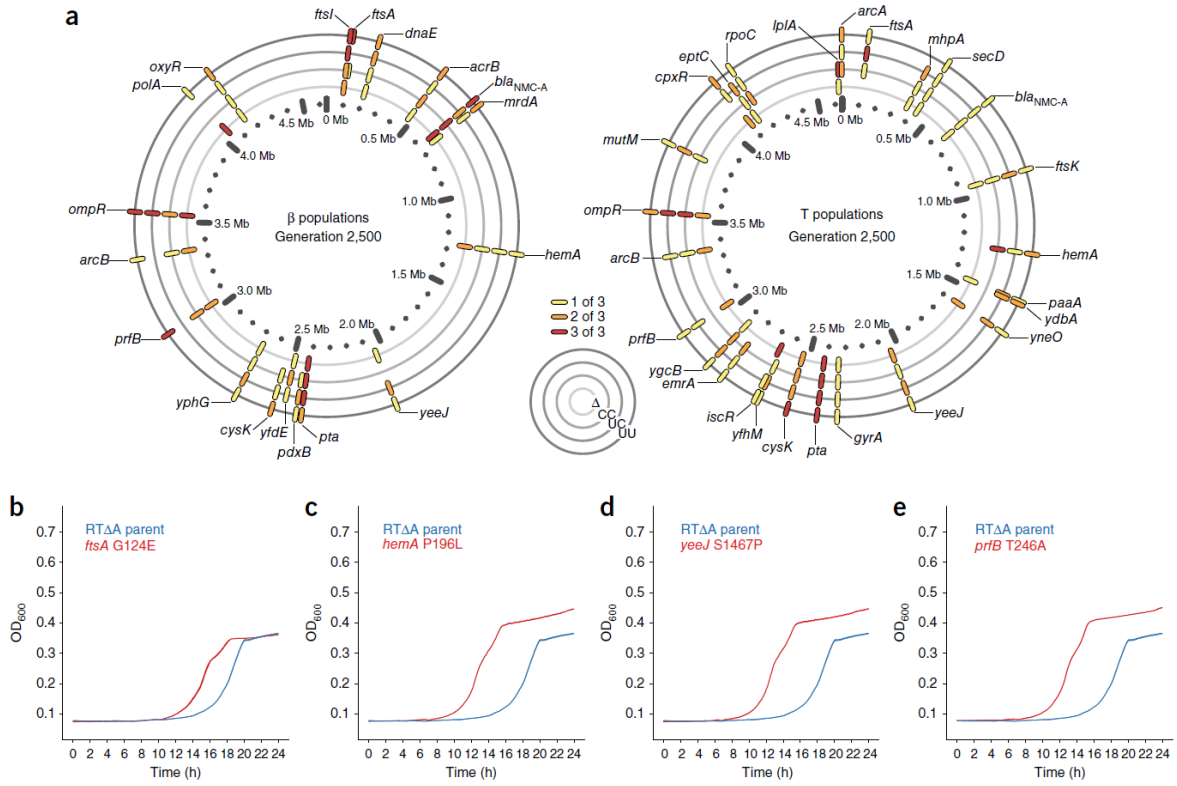


Figure 4.2 Genes mutated during continuous evolution.

(**a**) Mutations were identified by whole genome sequencing of all evolved populations after 2,500 generations. Genes that contained SNPs with >50% frequency in at least 4 of the 12 independent subpopulations are reported using a circular representation. From innermost to outermost, the rings represent the genome location with a scale in megabase pairs (Mb), the Δ populations, the CC populations, the UC populations, and the UU populations. The yellow, orange, and red bars represent one, two, or three mutant alleles in a given set of triplicate subpopulations, respectively. (**b**–**e**) Growth curves of *ftsA* G124E, *hemA* P196L, *yeeJ* S1467P, and *prfB* T246A mutants compared to the parental RTΔA strain. Curves are plotted with a line showing the mean and shading representing ± s.e.m., where n = 3 independent biological replicates. Note, *ftsA* G124E and *yeeJ* S1467P represent reversions to wild-type MG1655 sequence, which were not directly observed.

gene	β_Δ1	β_Δ2	β_Δ3	β_CC1	β_CC2	β_CC3	β_UC1	β_UC2	β_UC3	β_UU1	β_UU2	β_UU3
pta	A119T	W373R	W373R	W373R/G533D	P673S/V286A/A294T	S535P	A119V	S535P	H515R/W373R	**P673L**		**P673L**
ftsI		A257T	K483R/A537T	M471T	A228V/P311S		V545I	D76G	A537T	A537T	A3V/P311S	V545I
ompR	Y102C	Y102C	Y102C	Y102C		Y102C	Y102C	Y102C	Y102C	Y102C	Y102C	Y102C
prfB	T221A		N251D/T221A	T221A		K257R				T221A	T221A	T221A
acrB		M902I		L591P/A183T	F358L				A999T		A457T	L344P
dnaE		T470M			G925D		E916G	Y938H		R860H		E781G
yphG	A735V			E199G			S112F	T126A			P978L	
hemA	V183A	A173V		**P196L**			A173V					V183A
oxyR			C199R		V300A		M193I			A233T		L124P
yeeJ			A352T/G374S					V1184A	S1730N	G1888D		
pdxB				C220R				V212I	D162N		T299A	
arcB	V21A	Q39**			T652A							T655A
mrdA		A271T						T52A		A599T	Y520H	
ftsA				Q155R						T115A	G386R	A136V
yfdE			A308V	F53L	Y143C				V21I			
polA	S446F	T666A	M768V							N430S		
cysK				H153Y			I45N				G271D	T73A
bla $_{NMC-A}$	N216S	F105S/S2P/T217A	L169S/T218A	N216S/S2P	N216S	S2P		S2P/F105S	S2P/F105S	F105L	F105S/T217A	F105L

Table 4.2 SNPs enriched in populations evolved under β-lactam stress.

Red genes were mutated during construction of C321.dA and **red** SNPs directly reverse the mutation. The P673S SNP marked in **orange** indicates reversion to a non wild-type sequence.
A single asterisk indicates mutation to a TAG codon and a double asterisk indicates mutation to a TAA or TGA stop codon. TAG codon identity is dependent on the strain.

gene	T_Δ1	T_Δ2	T_Δ3	T_CC1	T_CC2	T_CC3	T_UC1	T_UC2	T_UC3	T_UU1	T_UU2	T_UU3
pta	359K/N519	S263N/**P673L**	W373R	S535P	S535P	T662I	D601A	A119T	**P673L**	W373R	E103K	G31D
ompR	R15C	Q30R/Y102C		Y102C	Y102C	Y102C	Y102C	Y102C	G236S/F29L	Y102C		Y102C
cysK	T73A		A281V/F144S	P68L		T69I/R100H	T73A	I45V		T73A	L284P	T73I
yeeJ	A704T		M1084I		D1233N		D1502E		E470K/A939T			V2034I/L1669F
ygcB			A522V	R389H/F257S		W406**/A142V	Q130*	R665C				I24L
ftsK			A97T		Q767*		Y217S		Q648*		Y638C/A870V	
ftsA						V112A	E201A/P98S	A294T	N145D		H159Y	
cpxR	D137N		V52A						H70Y/M76I	I4T	T220A	
hemA				I176N	A173V	**P196L**		P273L			A5V	**P196L**
arcA				T34I		G227S	P58S			L236P		D84G
mhpA	R392H			Q68*				K329E	P395L/R271C			
ydbA							R1001Q		A422T	T612A/G564D		K3R
iscR	T125A	Q44*	T125A						D88G	T4I		
yneO	V574I	T816A					L1142R		G707E	T87A		
paaA		Y94C					R138C		P24S	D248G		
rpoC				T1328A	D248G			R220H			A182V	
gyrA		V644A				A271T			H185Y			I590V
secD	S207G				A574V				P198S			E271G
eptC			V60A			Q248R		V155I	E374G			
yfhM						A1337V		G1122S	A787V			W450**
prfB		T221A	T221A						T221A		T221A	
mutM						N99S	P243S		W157**	Q212*		
emrA				V83A	A55T		L125W					W253R
arcB	Y234C	R271W			S507L			L386P				
lplA		V34I		A19V	A19V	F148L						
bla $_{NMC-A}$							S263A	F105S	T245A	T136A		

Table 4.3 SNPs enriched in populations evolved under thermal stress.

Red genes were mutated during construction of C321.dA and **red** SNPs directly reverse the mutation.
A single asterisk indicates mutation to a TAG codon and a double asterisk indicates mutation to a TAA or TGA stop codon. TAG codon identity is dependent on the strain.

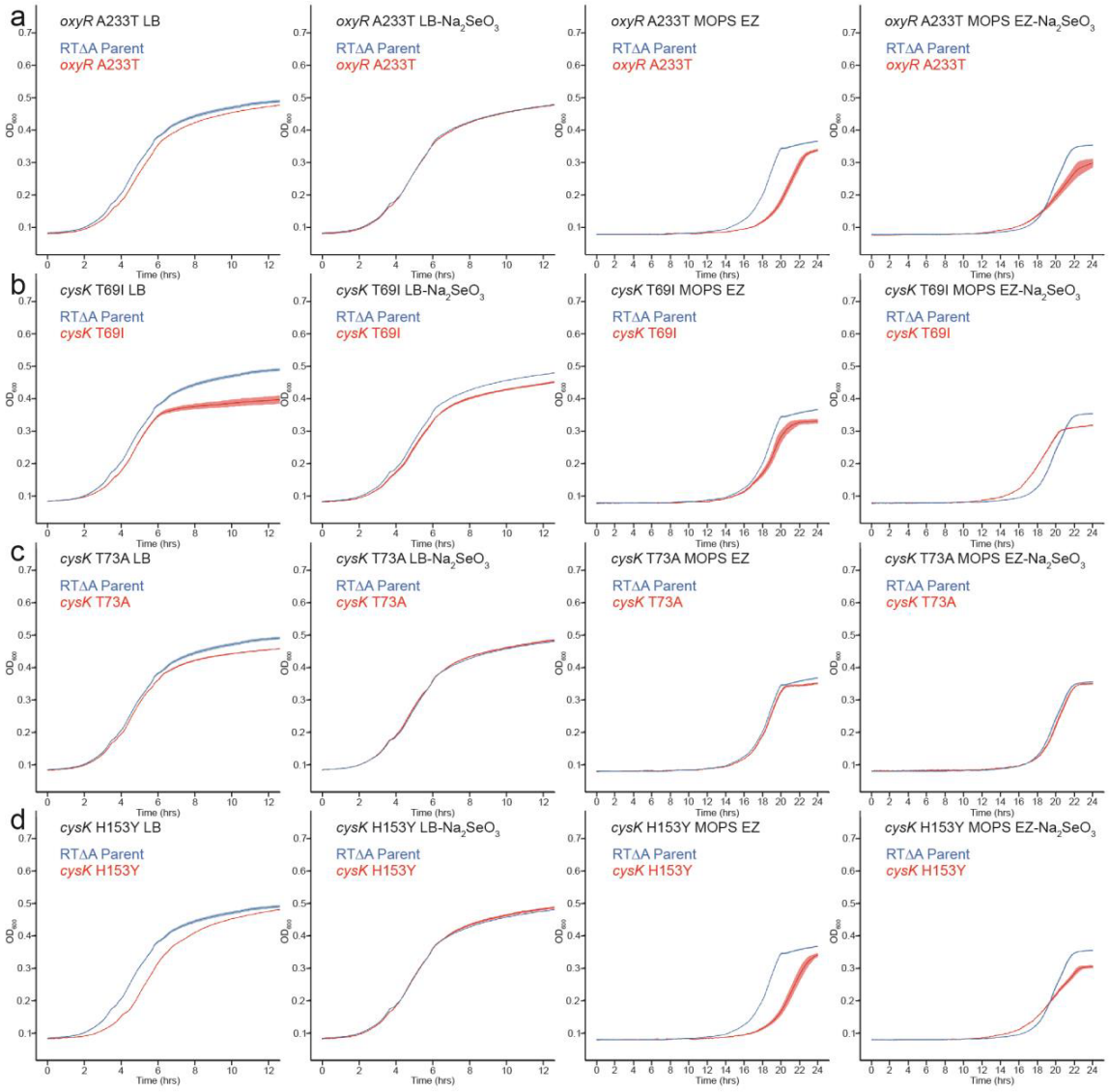


Figure 4.3 Growth curves of RTΔA cells containing point mutants in oxidative stress and selenite resistance.

Growth curves of RTΔA cells containing point mutants in oxidative stress and selenite resistance genes in rich and defined media ± 1 μM Na2SeO3. The curves are plotted with a line showing the mean and shading representing ± S.E.M. where n = three independent biological replicates. (**a**) Growth curves of RTΔA containing a constitutively active variant of OxyR (A233T) observed during evolution. The A233T mutation does not provide any clear benefits to cell fitness compared to a wild-type control. (**b-d**) Growth curves of RTΔA containing T69I, T73A or H153Y mutations respectively in CysK, which are expected to strongly inhibit cysteine biosynthesis, compared to a wild-type control. Mutations are not observed to have a significant impact on cell growth.

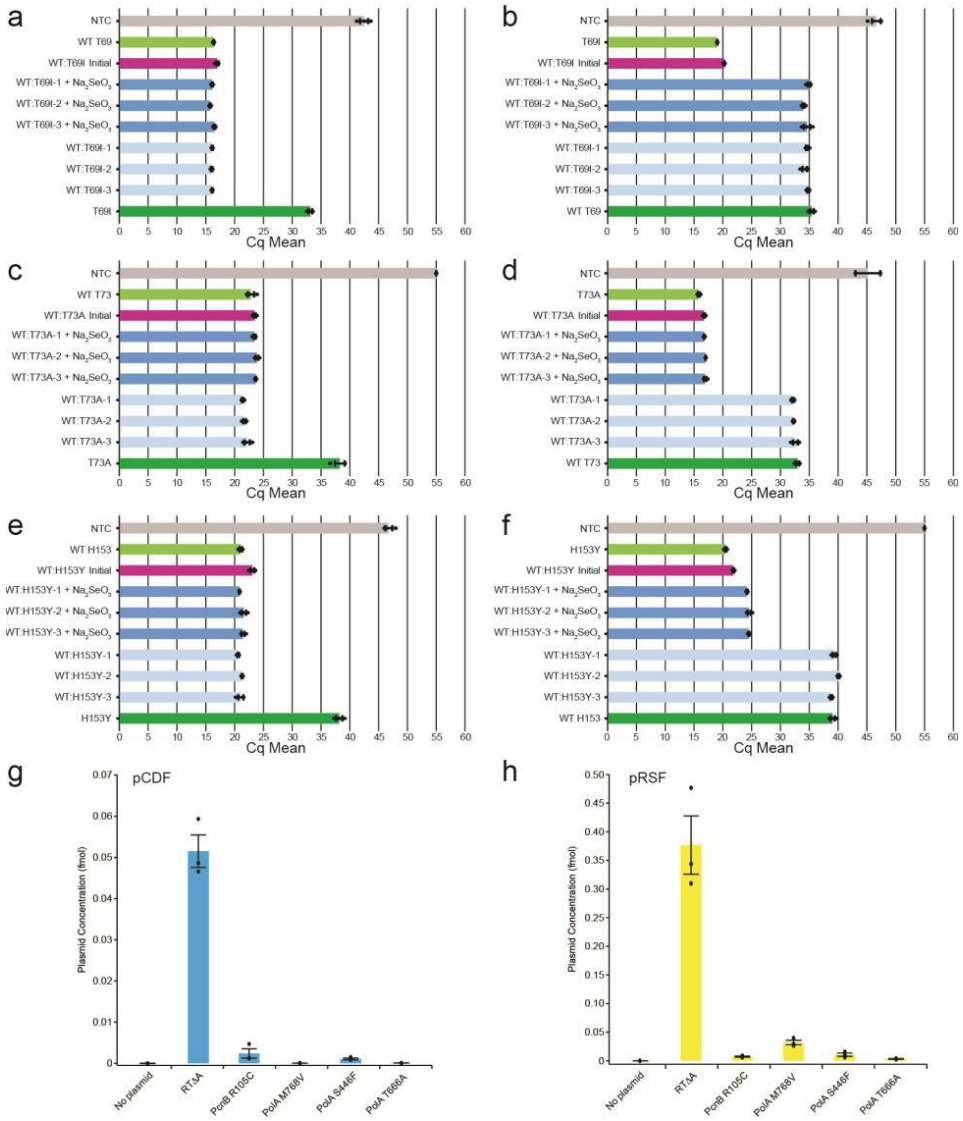


Figure 4.4 Mutant *cysK* allele retention and plasmid copy number determination.

Analysis of mutant *cysK* allele frequency by qPCR in cell populations passaged for 125 generations in the presence of 10 µM Na2SeO3. At generation 0 (purple bars), cell populations contained 50% wild-type *cysK* and 50% mutant *cysK*. (**a**) Detection of wild-type *cysK* (T69) in WT:T69I populations ± Na2SeO3. Wild-type *cysK* is detected at a similar frequency in both the presence and absence of Na2SeO3. (**b**) In contrast, mutant T69I undetectable (Cq = ~ 35) after 125 generations in all populations. (**c**) Detection of wild-type *cysK* (T73) in WT:T73A populations does not change ± Na2SeO3. (**d**) The *cysK* T73A mutant is lost from the populations when cultured in the absence of Na2SeO3 (light blue). In contrast, in the presence of Na2SeO3 (dark blue) the mutant allele is retained. (**e**) Detection of wild-type *cysK* (H153) in WT:H153Y populations is not affected by Na2SeO3 treatment. (**f**) Similar to T73A, the H135Y mutant is undetectable after serial passage in media without selenite (light blue), but is partially retained when selenite is present (dark blue). Abundance of pCDF (**g**) and pRSF (**h**) plasmids in RTΔA cells with mutations in poly(A) pol I (*pcnB*) and DNA pol I (*polA*). All mutations significantly decreased plasmid abundance compared to wild-type cells.

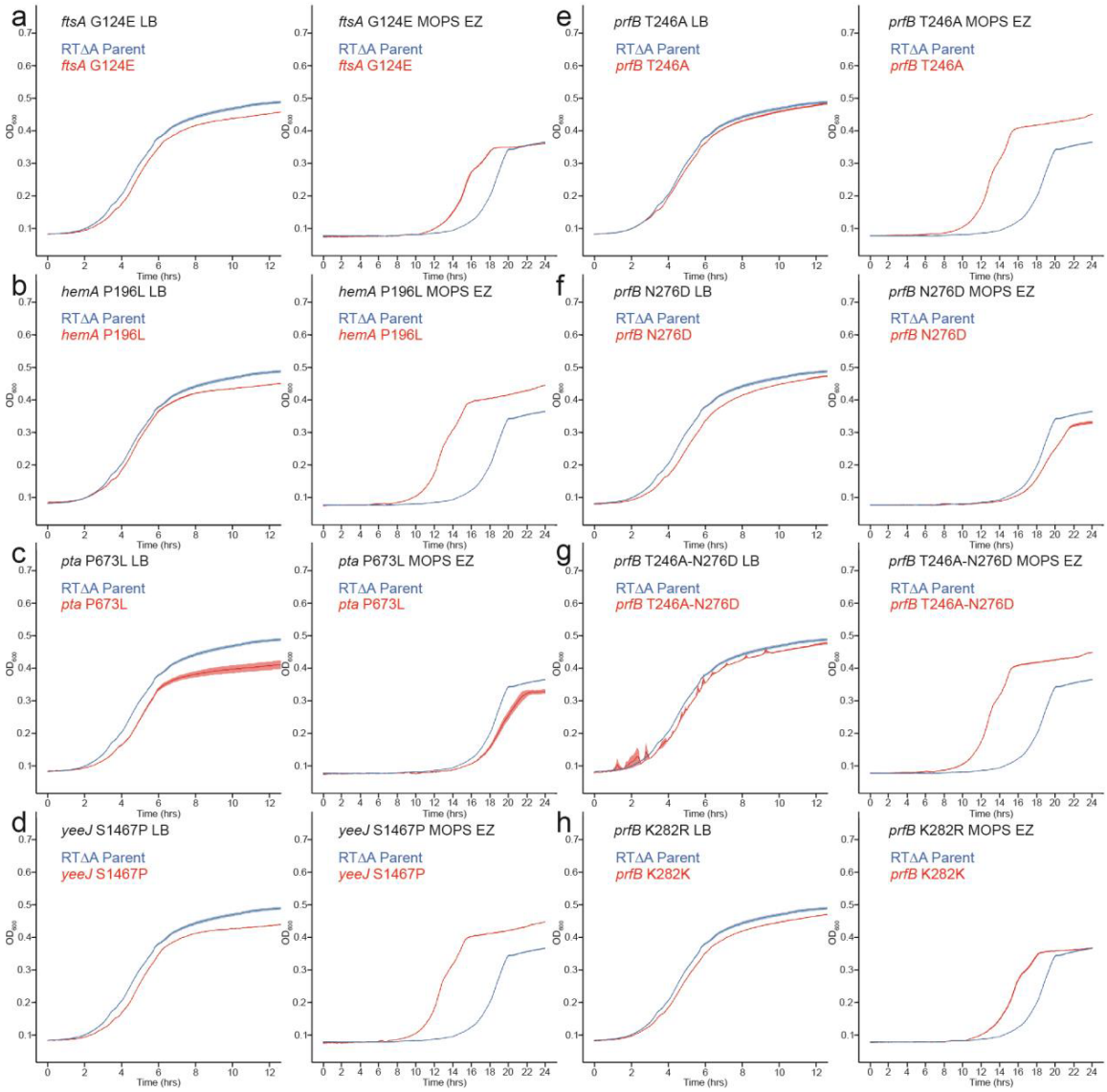


Figure 4.5 Growth curves of C321.ΔA and *prfB* (release factor 2) single point mutants in RTΔA cells in rich and defined media.

The curves are plotted with a line showing the mean and shading representing ± S.E.M. where n = three independent biological replicates. (**a**) Growth of RTΔA *ftsA* G124E shows minor improvement in defined media. (**b**) Growth of RTΔA *hemA* P196L shows significant improvement in defined media. (**c**) No improvement was observed in either rich or defined media for RTΔA *pta* P673L. (**d**) Growth of RTΔA *yeeJ* S1467P shows significant improvement in defined media. (**e**) The T246A mutation in release factor two significantly improves growth of RTΔA cells in defined media. (**f**) No improvement was observed in either rich or defined media for RTΔA *prfB* N276D. (**g**) A T246A-N276D double mutant was generated to investigate potential synergy between T246A and N276D which co-occurred in one evolved population. Observed growth improvement is due to T246A mutation only. (**h**) Growth of RTΔA *prfB* K282R shows minor improvement in defined media.

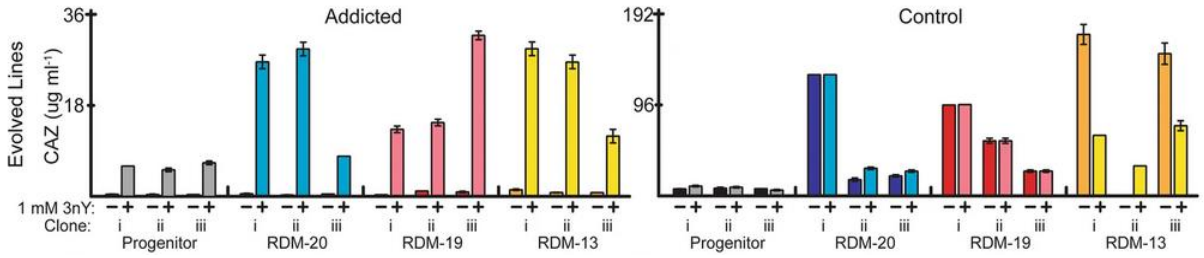


Figure 4.6 Ceftazidime MICs.

MICs of progenitor cells (gray and black) and evolved lines in the absence and presence of 1 mM 3nY. All lines increased MICs during evolution (upper panels, compare gray/black to colored bars). Lines addicted to 3nY remained dependent on 3nY for ceftazidime resistance after 2000 generations (upper left) while control lines never required 3nY for ceftazidime resistance (upper right). Plasmids extracted from evolved lines and transferred to wild-type *E. coli* strain MG1655 showed smaller increases in ceftazidime resistance (lower graphs). Values are the average of biological triplicates, error bars represent s.e.m.

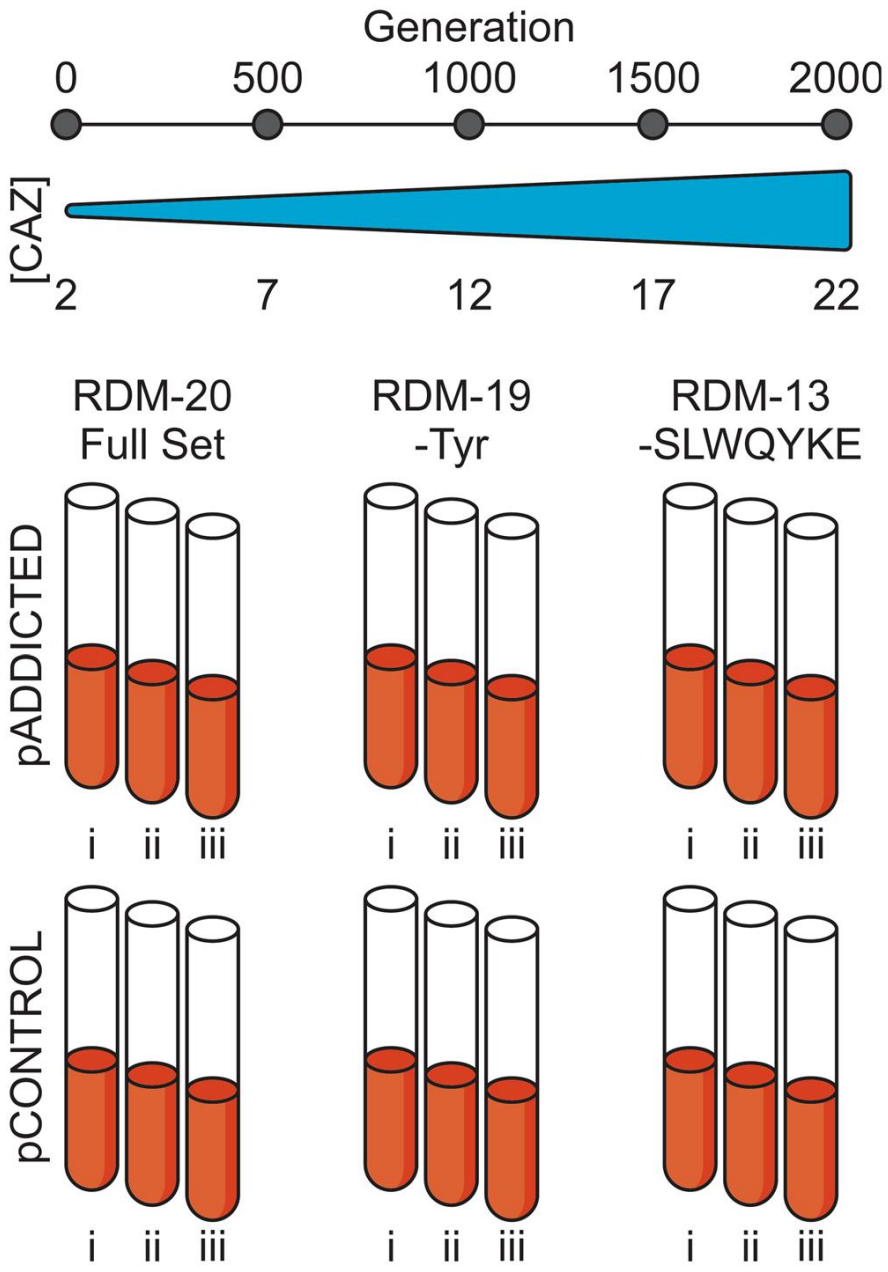


Figure 4.7 3nY evolution set up.

Wild-type *E. coli* strain MG1655 containing either pADDICTED or pCONTROL was evolved in biological triplicates (i, ii, and iii) for 2000 generations in one of three rich defined media conditions each supplemented with 10 mM 3nY. The first (RDM-20) contained all 20 canonical amino acids, the second (RDM-19) lacked tyrosine, and the third (RDM-13) lacked seven amino acids (serine, leucine, tryptophan, glutamine, tyrosine, lysine, and glutamate). During evolution ceftazidime concentration was increased to provide a fitness burden and enforce OTS activity.

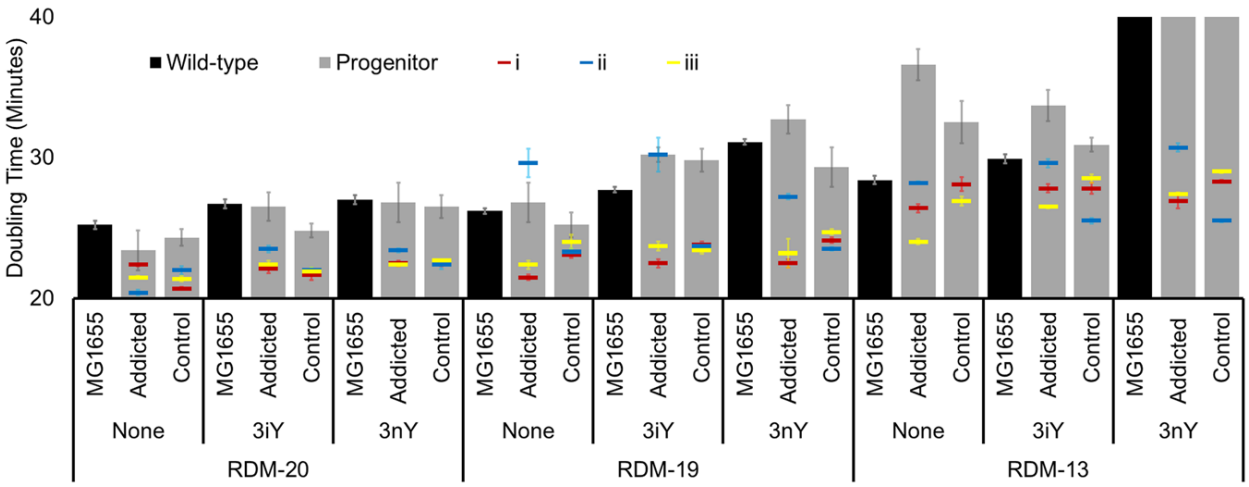


Figure 4.8 Growth rates of parental and evolved strains

Doubling time (in minutes) of wild-type *E. coli* strain MG1655 (black bars), and MG1655 containing the plasmid pADDICTED or pCONTROL (gray bars). Doubling times were measured in all three media conditions (RDM-20, 19, and 13) and with no ncAA (none), with 10 mM 3-iodo tyrosine (3iY), or with 10 mM 3-nitro-tyrosine (3nY) before evolution (gray bars). Doubling times of each evolved lineage (i, ii, and iii for each condition) were measure after 2000 generations of evolution (red, blue, yellow, respectively) in the media in which they were evolved, without ncAA, with 10 mM 3iY, or with 10 mM 3nY. Reported values are based on a minimum of two growth curves from biological replicates, error bars represent s.e.m.

Mutation	Lines	Effect	Previously Identfied
V33I	Addicted-13(ii)	—	[202,242]
Q39K	Control-20(ii)	Expanded Substrate Activity	[205]
G92D	Addicted-20(i), -20(ii), -19(ii)	Stabilizing	[202,243]
T139I	Addicted-20(iii), Control-19(ii)	*bla*TEM-189 specific	—
T140K	Addicted-19(ii), -19(iii), -13(iii)	—	—
M152I	Control-20(i)	*bla*TEM-189 specific	—
H153R	Addicted-20(ii)	Stabilizing	[85,244]
M155I	Control-20(iii)	—	[202,245,246]
M182T	Addicted-19(i)	Stabilizing	[75,82,205]
A184V	Addicted-13(i)	Expanded Substrate Activity	[247,248]
T200P	Addicted-20(i)	—	—
A224V	Control-19(i)	Stabilizing	[75]

Table 4.4 Mutations found in bla_{Addicted} and bla_{Control} in evolved lines.

Addicted				Control			
Line	Gene	Mutation	Position	Line	Gene	Mutation	Position
20(i)	sdaC	+9 insert (IS1)	925/1290	20(i)	sdaC	+9 insert (IS1)	1192–1200/1290
	yqiG	A > G	815/2439		waaQ	+9 insert (IS1)	581–589/1035
	envZ	M58R			ecpR/ykgL	C > T	311481
20(ii)	ydeT	A > C	1108/1227	20(ii)	opgG	+9 (IS1)	738–746/1656
	waaO	+9 insert (IS1)	531–539/1020		sdaC	**W18tag**	
	mtdL	A324T			waaO	+4 (IS5)	925–928/1020
	cyaA	−78	2420–2497/2547	20(iii)	sdaC	+1	798/1290
	dnaA/rpmH	−6	3883825		envZ	S87Y	
20(iii)	opgG	**Q314tag**			waaO	+5 (IS2)	100–104/1020
	clsA	+9 (IS1)	1307–1315		ytdD	A14G	
	sdaC	A383V		19(i)	ybbP	A580E	
	yghG	G63V			galU	H71L	
	envZ	F107C			uspC-tyrP	−11258 (IS1)	
	ykgM/ykgR	+9 (IS1)	313109		garD	K343Q	
19(i)	opgG	+9 insert (IS1)	385–393/1536		tusD	G43G	
	ycjO	−87	250–336/882		cyaA	D231G	
	waaO	+4 insert (IS5)	533–536/1020	19(ii)	opgH	H417D	
	ycaA	−1	2033/2547		uspC-tyrP	−11067	
19(ii)	yahF	K208R			waaB	+4 (IS5)	226–229/1080
	dosP	I282I			prpB/prpC	A > G	349642
	uspC-yecH	−10333 (IS1)			lacZ/lacI	2 bp > TT	366351
	lhgO	R346W			lacZ/lacI	C > T	366409
	yqeK	L30(tga)			lacI/mhpR	G > A	367573
	crp	F137L			ompF/asnS	G > A	987138
	tauA/tauB	G > T	386203		yhfA/crp	A > C	3485942
19(iii)	hofC	S27R		19(iii)	catC	I449S	
	uspC-yecA	−12537 (IS1)			quuD	+5 (IS2)	233–237/384
	ppx	−1	365/1542		tyrP	**W357tag**	
	rrlD	C > T	2209/2904		cyaA	D184N	
	envZ	G85V			ythY	R208L	
	dnaA	−54	251–304/1404	13(i)	ftsW	L141I	
	cyaA	R160L			galU	+12 (IS4)	695–706/909
	rrlD	C > T	3424575		uspC-yecA	−11714 (IS1)	
	ythU/ythV	G > T	4073632		tyrA	M117I	
13(i)	flgJ	+9 insert (IS1)	157–165/942		kduI	M193I	
	uspC-tyrP	−10976			mtr	−6	712–717/1245
	rpoD	P97L			malA/yfdC	G > T	2465297
	mtr	−6	712–717/1245		yhfA/crp	+5 bp (IS2)	3486024
	waaP	+4	413–416/798	13(ii)	lptD	**Q557tag**	
	cyaA	D231V			uspC-tyrP	−11031 (IS1)	
13(ii)	pykF	I264T			yhcG	S136A	
	uspC-tyrP	−11298 (IS1)			tnaA	+5 (IS2)	303–304/1416
	sanA	+9 (IS1)	165–173/720		ompF/asnS	+9 (IS1)	987064
	mtr	−6	391–396/1245		topA/cysB	+4 bp (IS4)	13337363
	ompR	M57L			mtr/dealD	T > C	3305872
	ytbA	+9 (IS1)	15–23/843	13(iii)	galU	+9 (IS1)	81–89/909
	waaQ/waaA	C-T	3808189		uspC-tyrP	−11280	
13(iii)	yafT	A > C	682/1458		mtr	L400R	
	ompF	+9 (IS1)	323–331/1089		mtr	−1	1011/1245
	cysB	−1	610/975		cyaA	C120Y	
	uspC-tyrP	−11018 (IS1)			rpsF	D13G	
	mtr	−6	712/717/1245		lrhA/alaA	A > C	2407094
	glpD	+9 (IS1)	1109–1117/1506				
	cyaA	P301L					
	ytiQ	K74taa					

Table 4.5 Genomic Mutations of Evolved Lines.

Genomic mutations which occurred after 2000 generations of evolution. Mutations which appeared after 125 generations in RDM-13 are underlined. The four in-frame amber codons which arose in ORFs are bold.

Chapter 5

Microbiota and Metatranscriptome Changes Accompanying the Onset of Gingivitis

The advent of next-generation sequencing has enabled better disease understanding at an unprecedented resolution. Genome scale mutations and global transcriptome quantification are now performed routinely, a stark contrast to the more traditional reductionist approach predominate as recently as a decade ago. One consequence of this increase in scale is that experiments can be more of a fishing expedition with no set expectation or outcome. That is, rather than requiring *a priori* hypothesis of the underlying molecular biology, deep sequencing provides a more comprehensive view and allows for the resulting data to guide the conclusions. Perhaps no medical subfield has benefitted more from a richer understanding at the nucleic acid level than the microbiome. The importance of bacterial communities is becoming more and more prevalent in not only maintaining homeostasis but also the role it plays in the onset of disease phenotypes. This chapter explores just that, analyzing the bacterial community changes that signal the onset of gingivitis.

Over half of adults experience gingivitis, a mild yet treatable form of periodontal disease caused by the overgrowth of oral microbes. Left untreated, gingivitis can progress to a more severe and irreversible disease, most commonly, chronic periodontitis. While periodontal diseases are associated with a shift in the oral microbiota composition, it remains unclear how this shift impacts microbiota function early in disease progression. We initiated this study in order to better characterize the progression from oral health to disease. We first analyzed changes in the abundances of specific microorganisms in dental plaque collected from teeth during health and gingivitis, the mildest form of periodontal disease. We found that clinical score of disease and patient from which the sample originated but not tooth brushing are significantly correlated with microbial community composition. While a

This chapter is adapted from Nowicki, E. M., Shroff, R., Singleton, J. A., Renaud, D. E., Wallace, D., Drury, J., ... & Scott, D. A. (2018). *mBio*, 9(2), e00575-18. I shared first authorship with EMN, and specific contributions are outlined in the text.

number of virulence-related gene transcripts are differentially expressed in gingivitis samples relative to health, not all are increased, suggesting that overall activity of the microbiota is dynamic during disease transition. Better understanding which microbes are present and their function during early periodontal disease can potentially lead to more targeted prophylactic approaches to prevent disease progression.

I contributed to this project by performing the bioinformatics analysis, data presentation, and writing of the manuscript.

5.1 Introduction

Oral microbes are found as an organized and complex polymicrobial biofilm community potentially containing at least 750 unique bacterial species with varying genetic potential[249–253]. While these microbes normally co-exist within the mouth as commensals, infrequent or inadequate cleaning can lead to periodontal disease in a susceptible host[253,254]. The mildest form of periodontal disease, gingivitis, is characterized by plaque buildup in the subgingival crevice of teeth[255] and inflammation of the gums[256,257]. Gingivitis symptoms can be eliminated and the gums restored to a healthy state through professional dental cleaning. Untreated gingivitis, however, can progress to chronic periodontitis[258], an irreversible periodontal disease characterized by chronic inflammation, destruction of gum tissue, and ultimately loss of both tooth attachment and alveolar bone[257,259] (**Figure 4.1**).

A number of studies have harnessed the power of next-generation sequencing technology to characterize both the composition and function of the oral microbiota during health or periodontitis[260–269]. A distinct phylogenetic structure in health relative to periodontitis has been revealed through 16S rRNA gene and shotgun metagenomic sequencing[260,265,267]. Other studies have performed metagenome[261,262,266,268] or metatranscriptome analyses[263,264,267,269] to examine the functional potential of the oral microbiota in health and disease. Several of these studies have revealed the presence and expression patterns of genes associated with pathogenesis during periodontal disease. Although the genomes of many commensals found during periodontal health contain virulence-related genes, many of these genes are either uniquely present or more highly expressed in samples collected from patients with chronic periodontitis[262]. Additionally,

several studies have revealed that gene expression patterns of the oral microbiota are patient-specific despite matched patient health status, suggesting that factors unique to each individual can shape microbial community structure and activity[262,264]. These studies have collectively expanded our understanding of how both community structure and function differ between periodontitis and health as well as across individuals.

Several other studies have focused on understanding changes in microbial community structure during health relative to early periodontal disease, gingivitis[254,256,257,270]. For example, Huang et al. reported that unique taxonomic groups are found in plaque samples from either healthy patients or those with gingivitis[270]. In another study, 50 adults with naturally occurring gingivitis were restored to "baseline" dental health and then subjected to a three-week experimental gingivitis treatment. Researchers found that 27 bacterial genera were differentially distributed between baseline and gingivitis, with 5 of these showing elevated abundance in health and 22 having elevated abundance in gingivitis[256]. While these studies have helped to elucidate the community-level shifts that occur during the transition from health to gingivitis, the functional changes that occur during the progression of periodontal disease have not yet been examined.

Understanding the functional contributions of the oral microbiota as periodontal disease develops is critical in order to develop more effective prophylactic treatments for preventing this all too common disease. The work herein assesses both community and functional changes of the human oral microbiota during the transition from health to an inflammatory periodontal disease. Analysis of subgingival plaque from a three-week experimental gingivitis treatment cohort revealed that similarity of microbiota composition between samples is significantly correlated with both clinical severity of gingivitis and patient, but not oral hygiene status. The relative abundances of seven of the most highly represented genera were found to differ significantly between patient-matched samples from teeth at different stages of gingivitis. With few exceptions, relative genus abundance as determined by 16S rRNA sequencing and relative transcript abundance determined by RNAseq were in agreement. Metatranscriptome sequencing of the plaque samples revealed that many genes significantly differentially expressed during gingivitis relative to health have virulence-related functions, and that while many of these functions serve to potentially promote tissue destruction and disease, the role of other virulence-related gene products

during early stages of disease is less clear. These data provide the characterization of changes in microbial activity that occur during the early stages of periodontal disease, which can potentially serve as targets to prevent further disease progression.

5.2 Results

5.2.1 Microbiota composition is correlated with clinical gingivitis index and patient

Subgingival plaque samples pooled from two brushed or unbrushed teeth in 10 individuals were collected at three different timepoints (as described in section 5.4.1), and were analyzed based on clinical disease parameters and community composition. Table 4.1 shows the day after study onset the sample was collected as well as two clinical measurements: the MGI (Modified Gingival Index) score, or clinical index of gingivitis severity of sampled teeth (depicted in **Figure 4.1**), and PD (probing depth), which measures the severity of tooth-gum attachment loss associated with periodontal diseases[271,272]. While MGI score increased over time in all teeth regardless of oral hygiene status (i.e. brushed or unbrushed), the magnitude of this increase was patient-specific. Although there were, on average, higher MGI scores for unbrushed teeth at each visit than for brushed teeth, this difference was not statistically significant (**Table 4.1**). Similarly, no significant difference in PD was found between brushed and unbrushed teeth between the start (visit 2) and end (visit 6) of the study (**Table 4.1**).

We first assessed the microbial composition and diversity within the collected plaque samples through 16*S* ribosomal RNA sequencing, allowing us to focus on organisms actively producing protein and thus with the greatest potential to influence community activity[273]. We performed alpha (within-sample) diversity analysis of sequenced 16*S* rRNA reads from subgingival plaque samples obtained from either brushed or unbrushed teeth. This revealed no statistically significant difference in Shannon index, which measures both species evenness and abundance, between samples originating from brushed or unbrushed teeth (**Figure 4.2a**). In contrast, samples grouped by clinical index of patient gingivitis severity (MGI score) showed significant differences in alpha diversity in both the high and medium MGI samples when compared to the low MGI group (**Figure 4.2b**, $p < 0.005$). While Shannon

index did not differ among most of the samples grouped by patient, samples collected from patients 6 and 14 did have higher diversity (**Figure 4.2c**).

Beta (between-sample) diversity was also assessed via a Bray-Curtis dissimilarity analysis of 16S rRNA reads and visualized using a Principal Coordinates Analysis (PCoA). Surprisingly, this analysis revealed no significant differences between samples collected from teeth with different oral hygiene status, suggesting that in our study brushing had no meaningful impact on the composition of the subgingival plaque microbiota (**Figure 4.3a**, p=0.685). Similarly, no significant phylogenetic similarity was found between samples when grouped by the net changes in pro-inflammatory cytokines IL-8 or MMP-8, and MMP-9 (**Figure 4.4**) between day 1 and day 21 of the experimental gingivitis study. Importantly, beta-diversity analysis revealed significantly different clustering of all collected samples by MGI score (**Figure 4.3b**, p=.001). This clustering was most distinct between samples collected from teeth with the lowest and highest MGI scores in the dataset (MGI=0 and MGI=2). Significant or marginally significant clustering by MGI score was also found when controlling for visit at which the sample was collected (**Figure 4.5**, visit 2 p=0.025, visit 3 p=0.012, visit 6 p=0.08) and for patient (**Figure 4.3c**, PERMANOVA, p=.001), though some study subjects did display tighter clustering than others (i.e. patients 6 and 14). Together, these data suggest that the strongest predictor of microbiota phylogenetic similarity between samples are the clinical index of disease severity (MGI score) and the patient from which the sample originated.

5.2.2 Shifts in relative genus abundance occur during health to gingivitis transition

In light of our beta-diversity results, patient-matched samples collected from teeth with the lowest MGI score (MGI=0; clinically healthy) or highest MGI score (MGI=2; clinical gingivitis) were assessed to determine the changes in relative abundance of microbial genera during disease progression. Patients 1, 3, 4, 5, and 15 had plaque samples collected from teeth that met this criterion (i.e. teeth with an MGI score of 0 and an MGI score of 2), and thus were included in this analysis; patients with samples collected from teeth that did not have an MGI score of 0 were excluded from this analysis. Samples collected from teeth with an MGI score of 0 at visit 3 were selected for further analysis due to increased amounts of nucleic acids in these samples relative to those from visit 2. Since our beta-diversity analysis

revealed no significant effect of brushing on sample clustering, data from samples with from both brushed and unbrushed teeth both with an MGI score of 0 were averaged when possible. Operational taxonomic units (OTUs) with low (<1%) relative abundance were filtered from the samples analyzed, and the remaining OTUs were pooled by genus. Clear shifts in the relative abundance of specific genera between health and gingivitis occurred within each patient analyzed (**Figure 4.6a**).

The mean count of genera in samples collected from teeth at times of health and gingivitis in these same five patients analyzed in **Figure 4.6a** was analyzed for statistically significant changes using a Kruskal-Wallis non-parametric ANOVA test, revealing 7 bacterial genera with significantly different abundance between samples at different stages of gingivitis in the same patients. Of these, *Streptococcus*, *Neisseria*, and *Lautropia* had significantly higher relative abundance in samples collected from healthy teeth (MGI=0, Visit 3), while *Oribacterium*, *Leptotrichia*, *Tannerella*, and *Lachnoanaerobaculum* were significantly more abundant during gingivitis (**Figure 4.6b**). This suggests a potential dysbiosis of microbial composition and increased abundance of periodontal pathogens occurs during the early transitional disease state.

5.2.3 Overall RNAseq transcript data correlates with 16S rRNA sequencing data

We next wanted to elucidate changes in the specific activities of the subgingival plaque communities during disease progression. As a proxy for functional activity, we used a metatranscriptome sequencing approach (RNAseq) to compare gene expression changes in plaque samples collected from teeth during a state of clinical gingivitis relative to health (MGI=2 vs. MGI=0 samples). Of the 5 study participants (patients 1, 3, 4, 5 and 15) from which samples were obtained from teeth with an MGI score of 0 at visit 3 and MGI score of 2 at visit 6, patients 3, 4 and 15 were selected for RNA sequencing. These three samples were chosen because they had higher RNA concentration and quality scores than samples from patients 1 and 5. Thus, we compared gene expression in samples collected from teeth during health (MGI=0) and gingivitis (MGI=2) that were both patient and timepoint (visit) matched. The total number of reads obtained after trimming and read mapping statistics and average coverage for each sample are shown in **Figure 4.7c**. Given the number of samples, minimum

average sequencing depth and effect size fold cutoff of 2.75, we calculated our statistical power to be greater than 0.8.

We first compared bacterial activity from transcriptomic data by plotting the within-sample normalized activities from both disease types after aggregating counts to common genera (**Figure 4.8a**). The most abundant genera in averaged plaque samples at MGI=0 or MGI=2 were identified and are indicated in **Figure 4.8a** by red points. In health, these genera were found to be *Streptococcus*, *Neisseria*, and *Capnocytophaga* with the former two having been identified in the 16S taxonomic analysis. We observed the genera favoring disease to be *Leptotrichia* (observed in 16S analysis) and *Prevotella*, and also *Fusobacterium*, for which metatranscriptomic influence in promoting gingivitis has been previously reported[274]. In addition to looking at changes in transcriptomic activity or read counts, which essentially measure the overall abundance of each genus in disease compared to health, we also analyzed the fold change of each genus during progression to periodontitis. While fold change is more sensitive to lower expressing genera, we can use this metric to characterize the largest relative community changes. We identified the most discordant genera in our transcriptomic analysis in diseased samples as *Tannerella*, *Treponema*, and *Leptotrichia* while *Haemophilus*, *Granulicatella*, and *Neisseria* were most discordant in healthy samples (**Figure 4.9a**). We next employed PCA to dissect the sample-level trends and observed well-defined clustering of the MGI=0 and MGI=2 patient samples (**Figure 4.8b**), thus showing distinct transcriptomic differences between samples during health and disease.

We next wanted to analyze the relationship between the taxonomic abundances and transcriptomic activity in samples collected from the same teeth during health or disease. To achieve this, we generated a rank-abundance plot to compare the bacterial community diversity (**Figure 4.8c**). Briefly, the counts to each genus were normalized by dividing by the most abundant/active genus within each sample and ordering by rank. We observe that, after averaging across three patients for each disease state (MGI=0 vs. MGI=2) and sequencing type, diseased patients exhibit greater bacterial diversity than healthy patients, and this trend is preserved through both the 16S and transcriptomic analyses. To more directly compare specific genera between abundance and activity, we performed regression analysis of average normalized genera in both MGI=0 and MGI=2 samples of the 25 most abundant genera (**Figure 4.10**). We find that microbial genus activity and abundance in

samples collected from teeth at a clinically healthy disease state are more correlated when compared to samples collected from teeth at a clinically diseased state. When analyzing the fold change of genera in healthy and disease samples, we find similar trends in the fold changes of each genus from disease relative to health in both the taxonomic and transcriptomic data (**Figure 4.9a**). Altogether, our data show that while the total abundance of certain genera differs when analyzed by either 16S rRNA sequencing or metatranscriptome sequencing, the overall trends in abundance fold change and rank demonstrate high concordance between the two datasets.

5.2.4 Virulence-related expression is elevated during the transition from health to gingivitis

We began our functional analysis by pooling read counts across all transcriptomes present in our samples by combining counts for transcripts with the same Enzyme Commission (E.C.) number. By pooling gene expression data for gene products involved in the same biochemical reaction, we were able to assess the overall activity of the microbial community. Further, previous work from our laboratory has shown that the expression of genes pooled by common metabolic function (E.C. number) is considerably less variable than the expression of individual organismal gene[264]. Differential gene expression between samples collected from the same teeth during health (MGI=0) and gingivitis (MGI=2) was analyzed using the R package DESeq2 [275]. Of the 2241 unique E.C. numbers analyzed (representing 111,778 of the 625,371 total unique ORFs with mapping reads in our reference dataset), 191 (8.5%) enzymes were significantly ($p < 0.05$) upregulated by 2.75-fold or greater in gingivitis relative to health while 180 (8.0%) were significantly downregulated. Overall, our data show that metabolic pathways more strongly associated with health (downregulated during disease) include genes involved in ascorbate and aldarate metabolism, porphyrin and chlorophyll metabolism, carbon-fixation in prokaryotes, the pentose phosphate pathway, antibiotic biosynthesis, and pyruvate metabolism. Metabolic pathways more strongly associated with disease (upregulated during disease) include genes involved in pyrimidine metabolism, vitamin B6 metabolism, glycolysis and gluconeogenesis, and propanoate and butanoate metabolism.

We then directed our attention to genes with virulence-related activities with significant changes in expression in gingivitis relative to health. We defined virulence-related gene

products as those involved in colonization, enhanced survival within, or evasion of host; or those that directly cause pathological damage associated with disease[276]. In addition to products commonly associated with virulence such as adhesins and antibiotic resistance genes, additional gene products found during periodontal disease that meet these criteria include those involved in bone resorption and tissue destruction[277]. 30 of the 191 significantly upregulated E.C. enzymes had virulence-related functions (**Figure 4.11**). Fold changes in expression for these genes within each patient and across all sequenced patients are shown in **Figure 4.11**, with peptidases, nucleases, and hydrolases shown in **Figure 4.11a**, and those involved in chemotaxis, cell surface modifications, and other virulence activities shown in **Figure 4.11b**.

The majority of non-specific peptidases were upregulated during disease transition, supporting the idea that expression of these potentially destructive enzymes can promote periodontal disease (**Figure 4.11a**). While some genes involved in the biosynthesis of cell surface features were significantly upregulated during gingivitis, others were significantly downregulated, including four genes involved in peptidoglycan biosynthesis. This suggests that changes to the cell surface or growth in general are dynamic during this early stage of disease. Genes involved in iron acquisition were similarly variably up or downregulated in our dataset. While *Vibrio*-specific siderophores were found to be downregulated during the early stages of periodontal disease, two other genes involved in iron acquisition were significantly upregulated across the entire metatranscriptome. Although the magnitude of gene expression changes varied among the three individuals sequenced as seen in other studies[262,264], collectively our data suggest that changes in the overall activity of oral microbiota during the early stages of periodontal disease progression promotes enhanced destruction of host tissue and survival within the oral cavity.

5.2.5 Individual periodontal pathogens upregulate expression of both specific and general virulence-related genes during gingivitis relative to health

Our next aim was to analyze the virulence-related activities of specific periodontal pathogens or opportunistic pathogens. We focused our analysis on representative species from the five most highly abundant genera present in our samples during disease (**Figure 4.8a**): *Leptotrichia* (*L. buccalis*), *Prevotella* (*P. nigrescens*), *Streptococcus* (*S. constellatus*),

Fusobacterium (*F. nucleatum*), and *Actinomyces* (*A. israelii*). We first normalized the total read counts for each transcript in the metatranscriptome across the three patients analyzed (patients 3, 4 and 15), pooling samples collected from the same teeth during health (MGI=0, visit 3) and during gingivitis (MGI=2, visit 6). We then analyzed differential gene expression between health (samples collected from teeth with MGI=0) and gingivitis (samples collected from teeth with MGI=2). This gave an overview of differentially expressed genes within the entire microbial community during gingivitis relative to health. We then looked at both known, characterized virulence-related genes associated with oral microbiota (collagenase, gingipain, hemagglutinin)[264] along with generalized virulence-related gene products known to promote bacterial colonization and survival in the oral cavity and to exacerbate periodontal disease (non-specific peptidases or proteases, stress response proteins). Criteria for virulence traits selected were the same as described in the overall metatranscriptome analysis in the previous section; virulence-related genes with the highest fold-changes in expression are shown. The differential expression data from the representative oral pathogens analyzed show a variety of virulence-related activities from these 5 organisms alone (**Table 4.2**).

L. buccalis virulence-related gene products upregulated during gingivitis include several genes involved in antibiotic resistance, non-specific proteases, as well as the response regulator MprA (**Table 4.2**). This gene has been shown in other organisms to play a role in regulation of genes crucial for pathogenesis[278], and thus may also play a role in *L. buccalis* pathogenicity. Both *P. nigrescens* and *F. nucleatum* significantly overexpressed a wide variety of virulence-related genes in plaque samples from teeth at a timepoint showing clinical indications of disease, including those involved in antibiotic resistance, proteolysis, breakdown of collagen, and iron uptake (**Table 4.2**). Of note, several gene products found to be increased in expression during gingivitis in our E.C. number analysis were found to be highly upregulated by *P. nigrescens*, including virulence factors endothelin-converting enzyme 1 and a gingipain. *S. constellatus* and *A. israelii* upregulated fewer genes across their entire genome and also fewer virulence-related genes compared to the other three organisms analyzed (**Table 4.2**). In addition to upregulating genes involved in general survival mechanisms or nucleic acid degradation, however, these organisms also upregulated gene products specifically shown to be involved in virulence in other organisms;

the Spx regulatory protein[279] (*S. constellatus*), and Virulence-associated protein 1[280] (A. israelii).

Upon analysis of the number of reads mapping to each of these species, we determined that upregulation of virulence-related genes in *P. nigrescens* was not necessarily induced, as an increase in both the number of active cells (as measured by rRNA sequences) and transcript abundance of this species occurred in teeth experiencing gingivitis (MGI=2) relative to the same teeth during health (MGI=0) (**Figure 4.9b, c**). Similarly *L. buccalis* 16S rRNA v4-v5 amplicon levels were also significantly elevated during gingivitis (**Figure 4.9b**), although transcript abundance of this species was not significantly different between the two disease states (**Figure 4.9c**). Despite the fact that increased expression of virulence-related genes is possibly due to an overall increase in the abundance of *P. nigrescens* and *L. buccalis*, it is still worth noting the changes in the overall functional repertoire of the oral metatranscriptome that can potentially promote periodontal disease progression; thus, we have included these genes in **Table 4.2**. On the other hand, relative 16S rRNA v4-v5 amplicon or transcript abundance of *S. contellatus*, *F. nucleatum*, and *A. israelii* did not significantly increase in samples collected from teeth with gingivitis relative to health (**Figure 4.9b, c**), suggesting that virulence-related gene products of these organisms are upregulated in their expression levels during disease as opposed to being elevated as a result of increased species abundance.

5.3 Discussion

In our analysis of dental plaque samples during the transition from health to gingivitis, we began by first analyzing community composition through a beta-diversity analysis. Composition and diversity of our samples was found to be significantly correlated with clinical index of disease severity (MGI score) (**Figure 4.3b**). Samples collected from teeth with no visible evidence of periodontal disease (MGI=0) cluster distinctly from those originating from teeth with the highest clinical disease score in our study (MGI=2). The clustering of our metatranscriptome sequencing data by MGI score (**Figure 4.8b**) along with our identification of differentially expressed virulence-related genes between samples from teeth with an MGI score of 0 vs. 2 lends support to the efficacy of MGI score as an indicator

of disease progression. We also found that samples originating from the same patient were statistically more similar in composition than those from different patients (**Figure 4.3c**), although samples still clustered by MGI score despite this patient effect. Our data do suggest that certain patients have more stable microbiota community structure while others were subject to greater community changes over the time course in which this study took place (**Figure 4.3c**). This implies that patient-specific factors not considered in this analysis, such as host genetics, co-morbidities, age, or gender could play an important role in shaping oral microbiota community structure.

Surprisingly, our beta-diversity analysis revealed that the oral hygiene status (i.e. whether or not teeth were brushed) had no significant impact on microbial community composition (**Figure 4.3a**). This is in contrast to at least one previously published study that found a significant decrease in plaque (as measured by gingival index) after brushing[256]. These patients, however, were subjected to a rigorous oral hygiene regimen to obtain this decrease. As it was not possible to ensure patient compliance with all study parameters during the course of the experimental gingivitis study, it is plausible that some patients did not adhere to the prescribed oral hygiene regimen. In agreement with this, Kistler and colleagues have published that many individuals do not self-apply oral hygiene techniques that are sufficient to prevent the onset and progression of gingivitis[257]. Although we were unable to accurately quantify the total raw number of bacteria present at any given time point (data not shown), it is also possible that brushing can affect the total number of microbes present rather than the diversity and relative abundances of different taxa. The net changes in levels of three different pro-inflammatory cytokines (IL-8, MMP-8 and MMP-9) measured in the gingival crevicular fluid of sampled teeth also had no significant contribution to microbial community composition (**Figure 4.5b, c, d**). As one report found that the levels of most inflammatory cytokines vary significantly between individuals[281], quantification of these particular cytokines may not be a reliable predictor of subgingival plaque community structure.

Our analysis of relative genus abundance within the subgingival plaque communities through both 16S rRNA sequencing and metatranscriptome sequencing revealed significant shifts during disease transition, although the two methods showed differences in the specific genera found to be most abundant (**Figures 4.6b and 4.8a**). For example, while both 16S

rRNA and metatranscriptome sequencing revealed *Streptococcus* and *Neisseria* to have higher abundance in samples collected from teeth at a point of clinical health, *Prevotella* was reported as a highly abundant genus during gingivitis based on metatranscriptome sequencing only (**Figure 4.8a**). This finding underscores the limitations of either method at determining genus abundance with total accuracy. Despite these differences, rank abundance analysis and ratios of composition (16S data) to activity (RNAseq data) of genera show similar trends in both datasets (**Figure 4.8c**, **Figure 4.10**). In addition to analyzing relative abundance in our plaque samples, we also analyzed fold-changes in genera (**Figure 4.9a**). The most discordant genera in total abundance between samples collected from teeth during health and gingivitis are not necessarily the highest in abundance (**Figure 4.8a**) according to our metatranscriptome data. While it would be interesting to further analyze specific functional contributions of these highly discordant genera during disease relative to health, limitations in total reads obtained made this difficult or impossible. For example, the lack of reads obtained for the genus *Tannerella* during health made further analysis of differential gene expression within this genus impractical, as virtually all genes show an increase in expression.

When analyzing our metatranscriptome data, we first wanted to get an overall idea of the general functional properties of the plaque communities during the transition from health to gingivitis. We thus began by pooling gene products (by E.C number) across all taxa and found that 8.5% of pooled gene products were significantly upregulated during gingivitis, while 8.0% were significantly downregulated. Genes involved in ascorbate and aldarate metabolism were upregulated during health, as seen in several previous studies[282,283]. Other metabolic pathways more strongly associated with health carbon-fixation in prokaryotes, the pentose phosphate pathway and pyruvate metabolism. Interestingly, porphyrin and chlorophyll metabolic gene products were also found to be associated with health. Genes involved in metabolic pathways including vitamin B6 metabolism and glycolysis and gluconeogenesis were found in our data analysis. As seen in other studies, metabolic pathways more strongly associated with disease were pyrimidine metabolism[265], and propanoate and butanoate metabolism[264,274,280].

We then focused our attention on significantly differentially expressed genes with virulence-related activity (**Figure 4.11**). These enzymes were involved in a variety of

activities including hydrolysis or proteolysis, degradation of nucleotides, chemotaxis, synthesis of cell surface structures (adhesion, motility, protection), and a variety of other virulence functions (**Figure 4.11**). Increased proteolytic and nucleolytic activity as seen in our data likely leads to the destruction of gum tissue and disturbance of the host immune system characteristic of periodontal disease, and have been noted as potential drivers of periodontal disease in other recent metatranscriptome studies[263,269,274]. Two different gingipains, or proteases that have previously been associated with periodontal disease in several studies[264,284–286], were significantly upregulated in our data. While general proteolytic, nucleolytic, and several chemotaxis related gene products showed increased expression, a number of other virulence-related gene products were downregulated in our data, suggesting the variable and unstable state of the microbiome during the early stages of gingivitis onset. Interestingly, several virulence-related gene products downregulated during disease relative to health may play a role in exacerbating the diseased state. For example, leukotriene A-4 hydrolase is typically involved in alleviating inflammation[287]; thus downregulation of this gene could lead to the promotion of an inflammatory state typical of periodontal disease. Downregulation of starvation-sensing protein RspA leads to an increase in the production of metabolites that signal starvation[288], and thus can lead to increased survival during nutrient limitation.

Finally, we wanted to analyze the contributions of specific pathogens to virulence. We chose to focus on representative pathogens or opportunistic pathogens from the 5 most abundant genera during clinical disease (MGI=2) (**Figure 4.8a**): *P. nigrescens* from *Prevotella*, *F. nucleatum* from *Fusobacterium*, *L, buccalis* from *Leptotrichia*, *S. constellatus* from *Streptococcus*, and *A. israelii* from *Actinomyces*. *P. nigrescens* and *F. nucleatum* were selected as representative pathogens because of their frequent presence in periodontal infections[289,290]. Other oral metatranscriptome studies have also shown that both of these organisms increase expression of virulence factors during periodontal disease[269,274]. From the genus *Leptotrichia*, we chose to look at *L. buccalis* as it is the most well-studied species of the genus and primary causative agent of opportunistic *Leptotrichia* oral infections[291,292]. Although many members of the *Streptococcus* genus are commonly occurring commensals, *S. constellatus* an opportunistic periodontal pathogen that is well-documented in periodontal disease, and thus our choice for a representative of the genus *Streptococcus*. Along with other

oral *Streptococci*, *S. constellatus* can also be synergistically pathogenic with *Porphyromonas gingivalis*[289,290]. Finally, although most members of the genus Actinomyces are associated with oral health, we selected Actinomyces israelii for further analysis because it is the causative agent of a rare but severe infection actinomycosis[293].

Within these 5 species, *L. buccalis*, *P. nigrescens*, and *F. nucleatum* overexpressed a variety of virulence-related genes and thus likely actively contribute toward disease progression, while *S. constellatus* and *A. israelii* overexpressed a more limited number of virulence genes during gingivitis (**Table 4.2**). Of note, *P. nigrescens* overexpressed a gingipain and several arginine/lysine-specific proteases during gingivitis (**Table 4.2**). In addition to degrading hemoglobin, these enzymes are able to degrade host cytokines and thus dampen the immune response to pathogens[294,295]. Interestingly, the metalloendopeptidase endothelin-converting enzyme was significantly overexpressed by *P. nigrescens* during gingivitis (**Table 4.2**). This enzyme can serve as a potent vasoconstrictor during cardiovascular disease[296], a common co-morbidity that can result following periodontal disease[297]. Other gene products with more generalized virulence-related activities that increased in expression among these organisms included peptidases, proteases, nucleases, and iron uptake proteins, gene products well-known to be involved in virulence in all organisms.

This study is unique in that we have assessed changes in both subgingival microbiota community structure and function associated with the transition from health to periodontal disease. Our findings regarding both the shifts in community structure as well as functional changes seen between health and gingivitis are largely corroborated by similar findings in previous work analyzing samples from healthy and diseased (periodontitis) teeth. These data therefore characterize how microbial activities change during the early stages of periodontal disease. Our hope is that the virulence-related genes identified here will serve as candidates in future studies for prophylactic targets that can potentially prevent progression from gingivitis to more severe forms of periodontal disease.

5.4 Methods

5.4.1 Patient population and study design

Samples were analyzed from a total of 10 patients that completed the present study. All patients were consulted and treated in the Graduate Periodontics Clinic at the University Of Louisville School Of Dentistry. The study was conducted in compliance with modern ethical guidelines, as approved by the local ethics board (Study # 14.0230). The study protocol was explained both verbally and in writing, and written informed consent was obtained from each participant prior to dental examination and sampling. The following were inclusion criteria in order for a patient to participate: 1) At least 18 years of age and in good general and oral health, 2) At least 20 natural teeth present, and 3) Baseline mean gingival index of less than or equal to 1 (according to the scoring by the Modified Gingival Index, or MGI). Conditions that would exclude an individual from participating were: 1) History of conditions requiring prophylactic antibiotic coverage prior to this study, 2) Use of antibiotic, anti-inflammatory, or anti-coagulant medication within one month prior to the study, 3) History of tobacco use, 4) Participation in another oral study involving oral care products concurrent or within 30 days of beginning this study, 5) Pregnancy or lactation, 6) Significant oral tissue pathology (excluding gingivitis), 7) Moderate to advanced chronic periodontitis or other form of periodontal disease, and 8) An underlying genetic or immunological condition that may influence the study (e.g. diabetes, immunodeficiency).

Patients that met all inclusion criteria during the first visit underwent a thorough periodontal examination by a trained dental hygienist. A professional dental cleaning was conducted, and instructions for proper oral hygiene were provided. Participants returned for a second visit (visit 2) within 2 weeks of the initial visit, at which time the study began. At visit 2, baseline clinical data was recorded, and both plaque and gingival crevicular fluid (GCF) samples were collected from the first molars. If the first molar was missing, samples were collected from the second molar. When patients brushed and flossed during the study, an acrylic stent was worn on either the top or bottom arch of teeth; these teeth were considered unbrushed (Figure 4.1). Whether the stent was worn on the top or bottom was randomly decided for each participant; half of the patients wore the stent on the top arch while half wore the stent on the bottom arch. The function of the stent was to prevent 'unbrushed' teeth from being cleaned during the study, and thus the stent was only worn during brushing and flossing. Patients were instructed not to use any other oral care products (such as gum, interdental cleaning aids, mouthwash, or chewing gum) during the

course of the study. After the baseline visit, subgingival plaque and GCF were collected two more times, at 3 days after visit 2 (visit 3) and up to ~3 weeks after visit 2 (visit 6). The day of the third sampling was determined by the amount of gingivitis progression; patients continued with the study until their MGI score was at or above 2.

5.4.2 Clinical assessment of gingivitis.

In addition to collecting plaque samples, calibrated examiners performed periodontal evaluations on the study participants. Examiners were calibrated to a gold standard examiner for probing depths (PD), plaque index (PI), and gingival index (GI). The examiners were calibrated until the agreement coefficient (Kappa statistic) is at least 0.90 for the PD, PI, and GI with the gold-standard examiner. Calibrated examiners scored patient teeth for experimental gingivitis using the Modified Gingival Index (MGI) on both the buccal and lingual marginal gingiva. Scores were given according to the following criteria: 0=Normal (absence of inflammation), 1=Mild inflammation of a single portion of the gingival unit (characterized by a slight change in color, but little change in texture), 2=Mild inflammation of the entire gingival unit, 3=Moderate inflammation (moderate glazing, redness, edema, and/or hypertrophy) of the gingival unit, 4=Severe inflammation (marked redness and edema/hypertrophy, spontaneous bleeding, ulceration of the gingival unit). Probing depths were measured (in mm) at a total of six sites per tooth on both the buccal and lingual marginal gingiva, and average measurements were recorded.

Gingival crevicular fluid sampling (GCF) samples were collected from 8 sites by inserting paper strips into the sulcus of each site for a total of 30 seconds. GCF volume was measured using an electronic measuring device. Subgingival plaque samples from two brushed or two unbrushed teeth from each individual were collected at the three timepoints after the onset of the experimental gingivitis study. The two brushed or unbrushed plaque samples at each timepoint were pooled prior to nucleic acid extraction and then immediately frozen at -20 °C until further sample processing. Subgingival plaque samples were collected from the same sites during the course of the study by inserting one sterile endodontic paper point (Dentsply Caulk, Milford, DE) into the sulcus of each tooth for 10 seconds, followed by scraping with a curet. Upon collection of subgingival plaque, samples were immediately placed into a

microcentrifuge tube containing RNALater (Invitrogen) and stored at -20 °C until further sample processing. Samples from two brushed or two unbrushed teeth were pooled to ensure that enough microbial cells were collected in order to have sufficient amounts of DNA and RNA for downstream analyses, and also to account for the inherent differences seen from tooth to tooth within the same individual.

5.4.3 Cytokine analysis.

Paper strips used to collect GCF were thawed on ice and GCF was eluted by adding 200 □L elution buffer (50 nM Tris-HCl + 5 mM CaCl2, 0.2 M NaCl (pH 7.6), 1 mg/L antipain, 1 mg/L aprotinin, 125 mg N-ethylmaleimide, and 50 mg detergent (Zwittergent 3-12, EMD Millipore)). Elution was achieved by vigorous vortexing for 15-minute intervals for a total of 1 hour. Cytokine analysis was performed using a commercially available multiplexed bead-based assay designed to quantitate multiple cytokines. IL-8, MMP-8, and MMP-9 were measured using Luminex technology in pooled GCF samples.

5.4.4 RNA isolation and preparation.

Total RNA was prepared from frozen subgingival plaque samples stored in RNALater (Invitrogen) as described previously [22]. One half of the purified RNA was used for rRNA sequencing of the v4-v5 region of 16S rRNA cDNA, while the other half of the sample was saved for subsequent treatment with RiboZero and RNA-seq library preparation.

Pro- and eukaryotic ribosomal RNAs were removed from total RNA using the RiboZero Epidemiology kit (Epicentre). mRNA was purified, fragmented, and precipitated as previously described[264]. RNA-seq libraries were constructed using the NEB Next Multiplex Small RNA Library Prep Set for Illumina following the manufacturer's protocol. The cDNA libraries obtained at the end of this prep were stained using SYBR gold nucleic acid stain (Invitrogen) and visualized using a GBox imaging system. cDNA between ~150 and 300 bp was extracted according to the NEB QC Check and Size Selection protocol (protocol E7300), and resuspended in RNase-free water. Library cDNA concentration was determined using a Qubit fluorometer (Thermo-Fisher), and size distribution was analyzed on an Agilent

Bioanalyzer. Single-end 50-bp sequencing was performed at the UTGSAF on an Illumina HiSeq2000 system.

rRNA sequencing of the v4-v5 region of 16S rRNA was performed as previously described[264]. Briefly, total RNA was reverse-transcribed into cDNA using the random primer 16S926RT. The ~500bp v4-v5 region of the 16S rRNA gene was then PCR-amplified from cDNA using primers 16SV4515F and one of the 24 uniquely barcoded reverse primers. Custom primers were then used for sequencing on the Illumina MiSeq platform. All samples have been uploaded to the NCBI Sequence Read Archive (SRA) under BioProject ID PRJNA387475.

5.4.5 Bioinformatic analysis

The 16S rRNA V4/V5 MiSeq sequencing reads were assembled, mapped and analyzed using python scripts within the open-source phylogenetic DNA analysis pipeline, Qiime[298]. Paired-end reads were assembled and then quality filtered using the Qiime python scripts multiple_join_paired_ends.py and multiple_split_libraries_fastq.py. Reads were mapped to the HOMD 16S rRNA reference (v14.51) to determine the number of reads mapping to each distinct OTU using the Qiime python script pick_closed_reference_otus.py. Relative abundances of each OTU were then summarized and visualized using the Qiime python scripts summarize_taxa.py and summarize_taxa_through_plots.py. Sequencing reads were rarefied 10 times at each step with a step size of 500-sequences from 500 to 5,000 sequences using the Qiime python script multiple_rarefactions.py. Mean alpha diversity (within-sample diversity) was calculated using the Qiime python scripts alpha_diversity.py and collate_alpha.py. These scripts were used to determine the Shannon index for each sample or group of samples, which is a measure of species richness within samples. Beta diversity (between sample) analysis was performed by running the Qiime python script beta_diversity.py using Bray-Curtis dissimilarity analysis as a measure of beta diversity. Samples were then clustered by principal coordinate analysis using the Qiime python script principal_coordinates.py. Samples were categorized, and significant differences between categories were determined by PERMANOVA statistical analysis using the Qiime python script compare_categories.py. For analysis of relative genus abundance in samples, OTUs not present in at least 2 samples or less than 1% abundance in any sample were filtered from

the data using the Qiime python script filter_otus_from_otu_table.py. Relative abundance data was then calculated using the summarize_taxa_through_plots.py script. To calculate genera with significant changes in abundance between sample groups, relative abundances were first converted to absolute read counts using the summarize_taxa.py script with the –a flag. The Qiime python script group_significance.py was the used to calculate significant changes, using the Kruskal-Wallis analysis as the significance test.

Following RNA-seq, HiSeq reads were downloaded, concatenated and adapters were removed using Flexbar (v2.34)[299], as described previously[264]. Raw sequencing reads have been uploaded to the NCBI Sequence Read Archive (SRA) under BioProject ID PRJNA387475. Reads were trimmed to a minimum length of 18-bp or 30-bp, depending on the analysis. An absolute minimum length of 18-bp was selected for the pooled reads to minimize the likelihood of random, non-specific mapping to our reference metagenome. Since these functional data are pooled across all species by E.C. number and did not aim to detect specific species-level reads, this amount of random mapping would affect our results only minimally (Figure 4.7B). For species-specific functional analysis, 30-bp trimmed reads were used to obtain a minimum of 10X coverage of the reference metagenome (Fig. 4.7A). For samples lacking enough 30-bp trimmed reads to obtain 10X coverage, reads between 18 and 30bp were used to ensure adequate coverage of the search space.

The reference metagenome was generated as previously described[264]. All annotated human oral microbiome genome sequences and annotations were downloaded in both Fasta and GFF formats from the open-access, online Human Oral Microbiome Database (HOMD)[300]. Genome sequences were downloaded, concatenated, and processed to remove non-protein coding genes using the Perl scripts GenomeMerge.pl and HOMDpull.sh to create an annotated metagenome to map our sequencing reads to. If an E.C. number was associated with a gene, it was downloaded from KEGG[301] using the custom Perl scripts PullEC.pl and HOMD_GenomeMerge.pl. Custom scripts are available at the following web address: http://github.com/khturner/metaRNA-seq.

Trimmed RNA-seq reads were mapped against the reference metagenome using a Bowtie2 end-to-end alignment, very-sensitive parameters[302]. The species read match with the highest MAPQ score for each read was kept; other potential matches of lower quality score were discarded. In the event of multiple matches with equal quality score, bowtie2

selected the match to be kept at random. After mapping with bowtie2, reads with >2 mismatches and unmapped reads were discarded using the UNIX grep command. To compare samples across the same condition, within-sample normalization was performed by dividing the counts to each genera by the most highly activity genus for that sample. The statistical power of our study, and thus the likelihood of identifying a true positive, was found to be 0.8 when the effect size is set at 2.75 fold cutoff, calculated using the R program RNASeqPower (v3.5)[303] within the Bioconductor package using a sample size of 3, a false discovery rate of 0.05, and the lowest sequencing depth in all libraries (determined through samtools mpileup).

After removing unmapped or highly mismatched reads and sorting the remaining reads, the number of reads mapping to each ORF was determined using the open source python library, HTSeq[304]. To analyze reads corresponding to specific species-level ORFs, genus, species and product data were overlaid with the count matrix resulting from HTSeq using UNIX commands. To analyze reads pooled across all organisms by E.C. number, E.C. numbers corresponding to ORFs were overlaid with the count matrix, and genes without an E.C. number were removed from the list using UNIX commands. Counts for ORFs from different species but corresponding to the same E.C. number were then combined using the data.groupby command within the open source python library, pandas. For both species-specific ORF read counts and pooled E.C. number counts, count tables were imported into Rstudio. Differential expression was then assessed using the open source R package, DESeq2[275]. Low vs. high MGI score samples were analyzed in a paired manner, to account for differences between patients. Differential gene expression across all patients was analyzed by grouping the individual patient read counts into MGI=0 or MGI=2 categories in DESeq2, treating each patient as a replicate in each category. Differential gene expression for individual patients was calculated from patient raw read counts normalized using DESeq2; read counts for specific E.C. numbers analyzed obtained for MGI=2 samples were divided by counts for MGI=0 samples. For samples with <1 read for a particular gene, a minimum read count of 1 was used in order to calculate fold changes.

5.5 Conclusions

While the other chapters in this work provide insight into how biology can be leveraged to engineer improved or novel properties, the study presented here differs in that a top-down approach is used to better understand disease progression. We narrow our focus to analyze the microbiota changes in both oral community abundance and transcript activity in the progression towards gingivitis. The study was conducted with a unique design, as healthy and disease were stratified within each patient, greatly reducing the person-to-person variation in the oral microbiome composition. Enabled by high throughput sequencing, interest in the microbiome has prompted numerous similar studies at different stages of disease severity, and our results are largely corroborative. Repetition and external confirmation are necessary to develop a more robust view of periodontal disease progression. Hopefully, the virulence-related genes identified here will serve as candidates in future studies for prophylactic targets that can potentially prevent progression from gingivitis to more severe forms of periodontal disease

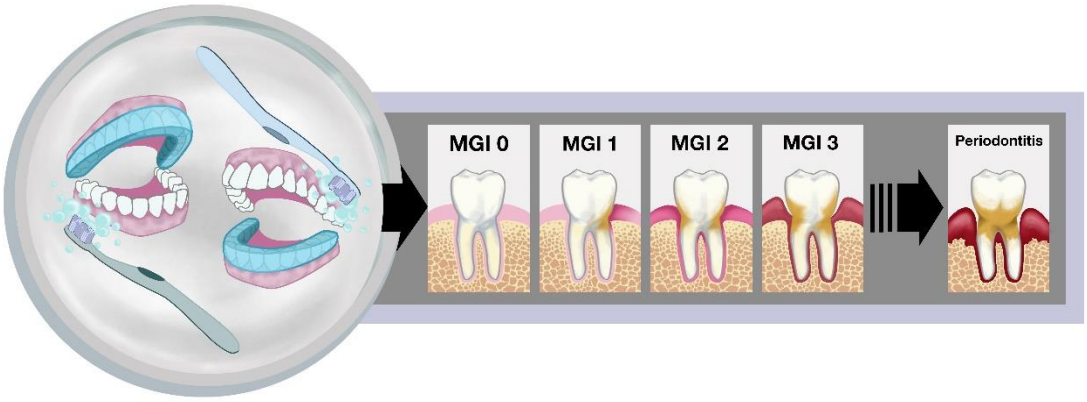


Figure 5.1 Study design and visualization of the progression from health to periodontal disease.

On the left, the covered or uncovered teeth depict the study design utilized, in which an acrylic stent (shown in blue) was worn to cover either the entire top or bottom set of teeth during brushing throughout the course of the treatment. The images on the right illustrate the clinical symptoms associated with gingivitis that were scored by trained dental professionals in this study. An MGI score of 0 represents a healthy tooth with no indication of inflammation (shown by increasing redness at the gum) or plaque (tan color on tooth). Healthy teeth progress through varying degrees of gingivitis as depicted by the MGI 1, MGI 2, and MGI 3 panels, and can eventually progress to the chronic gum disease periodontitis shown on the far left.

Patient:	Gender:	Age:	Oral hygiene status:	Visit 2:				Visit 3:		Visit 6:			
				Study Day:	MGI:	Buccal PD:	Lingual PD:	Study Day:	MGI:	Study Day:	MGI:	Buccal PD:	Lingual PD:
1	M	28	Brushed	1	0	2.67	2.67	3	0	21	1	3	2.67
			Unbrushed		0	2.33	3		0		2	2.83	3.5
2	F	30	Brushed	1	1	3	3.17	3	1	21	1	2.67	3.17
			Unbrushed		1	3	3.17		1		2	3	3
3	M	31	Brushed	1	0.5	2.17	3	3	0	21	1	1.83	3
			Unbrushed		0.5	2	2.67		0		2	2.17	3
4	F	53	Brushed	1	1	3	3.17	3	0	21	1.5	3.33	4.17
			Unbrushed		1	3	3.33		1		2	3.67	4
5	F	25	Brushed	1	0	2	1.83	3	0	22	1	2.5	2.67
			Unbrushed		0	1.5	1.83		0		2	2.67	2.5
6	M	24	Brushed	1	1	2.83	2.83	4	1	25	2	2.67	2.67
			Unbrushed		1	2.67	2.83		1		2	3.17	2.83
11	F	19	Brushed	1	1	2	1.83	4	1	26	2	2.5	2.33
			Unbrushed		1	1.83	2.83		1		2	2.5	3.33
13	F	20	Brushed	1	1	2.83	2.83	4	1	17	2	3	2.83
			Unbrushed		1	2.67	2.83		1		2	2.67	3
14	M	23	Brushed	1	1	3	3.5	6	N/A	25	2	2.83	3.33
			Unbrushed		1	3	2.67		N/A		2	4	2.67
15	F	21	Brushed	1	0.5	2.67	2.83	3	0	21	1	2.5	3.17
			Unbrushed		1	2.5	2.83		1		2	3	3
Brushed average:				Visit 2	0.700± 0.126	2.617± 0.129	2.766± 0.173	Visit 3	0.444± 0.176	Visit 6:	1.450± 0.157	2.683± 0.128	3.001± 0.161
Unbrushed average:					0.750± 0.134	2.450± 0.167	2.799± 0.126		0.667± 0.167		2 ± 0	2.968± 0.171	3.083± 0.137

Table 5.1 Patient clinical plaque sample data.

Patient number and demographics, oral hygiene status (brushed or not brushed) of teeth, MGI score (Modified Gingival Index), and PD (Probing depth, mm) are indicated. MGI scores progressed from healthy (0) to mild gingivitis (2)—see Materials and Methods (and Figure 1) for an expanded definition of MGI score. MGI and PD measurements represent the average of 2 sampled teeth for each condition. The average measurements and standard errors of the mean for brushed and unbrushed samples across all patients at each visit are indicated. Student's T-test revealed no significant differences in MGI score and PD between brushed and unbrushed teeth.

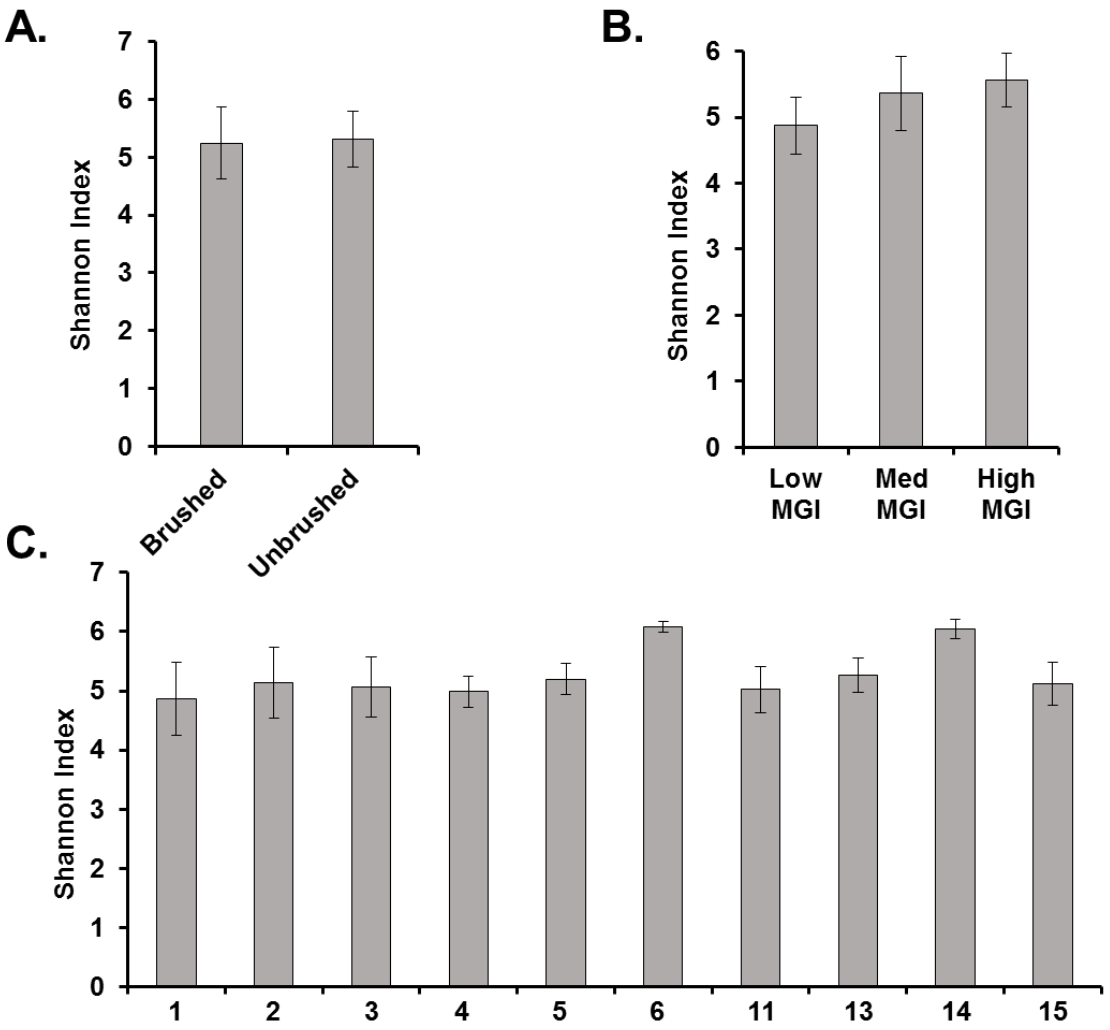


Figure 5.2 Shannon index of within sample (alpha-diversity) of 16S rRNA sequencing reads from all samples.

Samples are grouped by A) Oral hygiene status of the tooth (i.e. whether or not tooth was brushed), B) MGI score (Low MGI=0-0.5, Med MGI=1-1.5, High MGI=2; both Med and High MGI were significantly higher than Low MGI, $p < 0.005$) and C) Patient. Error bars show standard deviation in samples.

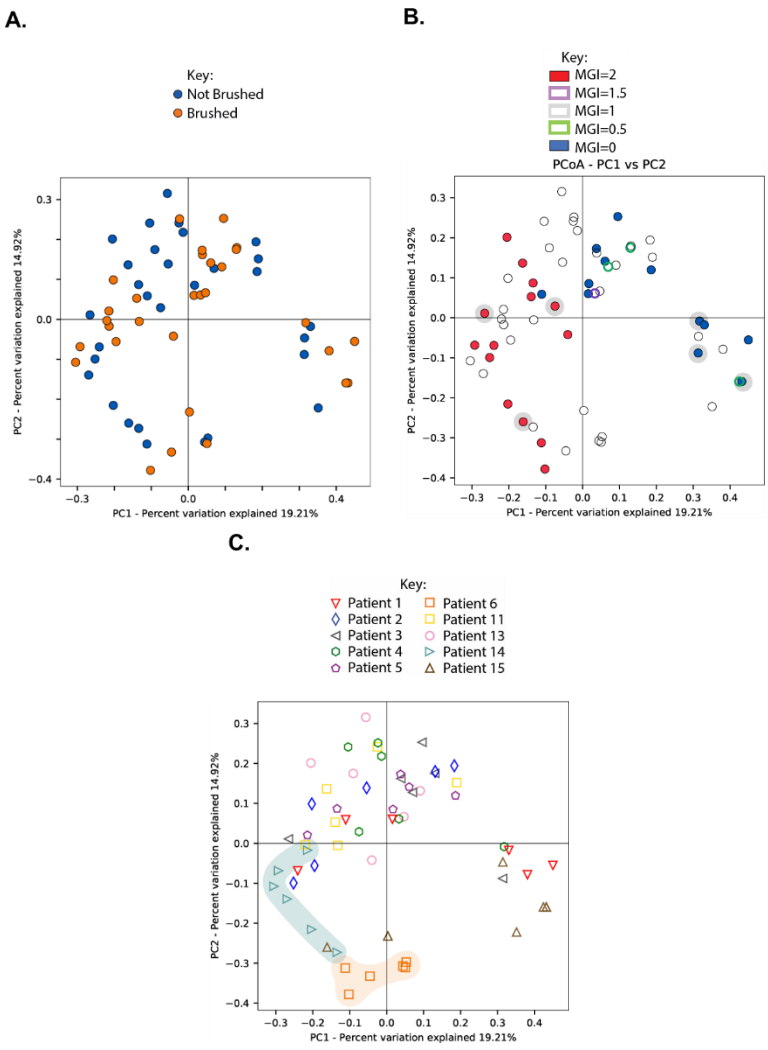


Figure 5.3 PCoA analysis of subgingival plaque sample 16S rRNA sequencing reads.

16S rRNA sequencing reads were categorized into distinct operational taxonomical units (OTUs) by mapping to the HOMD (v14.51) Reference database using standard Qiime scripts. Beta diversity between samples was measured through a Bray-Curtis dissimilarity analysis, and the principle coordinates (PCo) are plotted and colored by A) Brushed/not brushed plaque samples (p=0.68); B) MGI score (p=0.001) (samples chosen for RNAseq are highlighted); and C) Patient of origin (highly clustered patients are highlighted) (p =0.001).

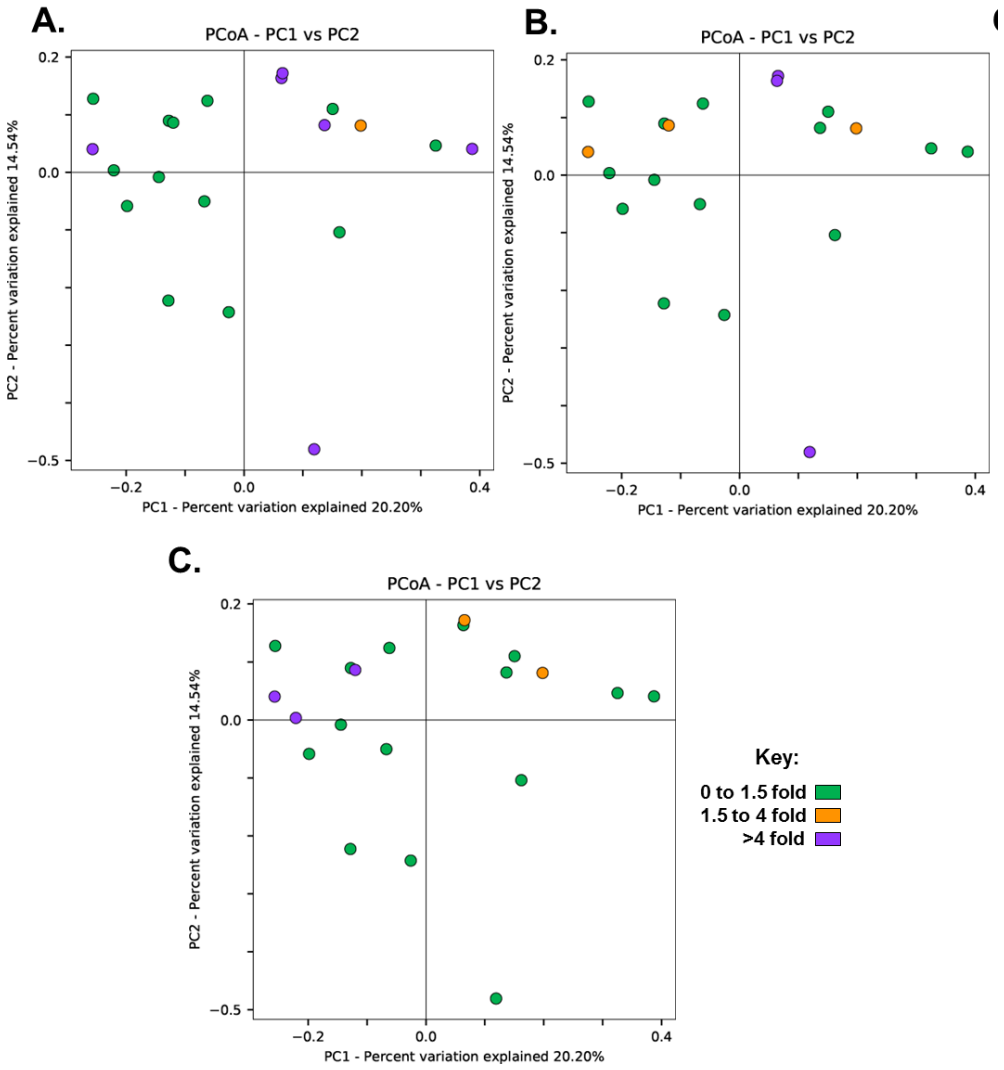


Figure 5.4 PCoA clustering of 16S rRNA sequencing samples by changes in cytokine levels.

16S rRNA sequencing reads were categorized into distinct operational taxonomical units (OTUs) by mapping to the HOMD (v14.51) Reference database using standard Qiime scripts. Beta-diversity between samples was measured through a Bray-Curtis dissimilarity analysis of the different OTUs within samples, and the principle coordinates plotted and colored by: A) Net fold change in IL-8 concentration during study, p=0.347; B) Net fold change in MMP-8 concentration during study, p=0.831; C) Net fold change in MMP-9 concentration during study, p=0.067.

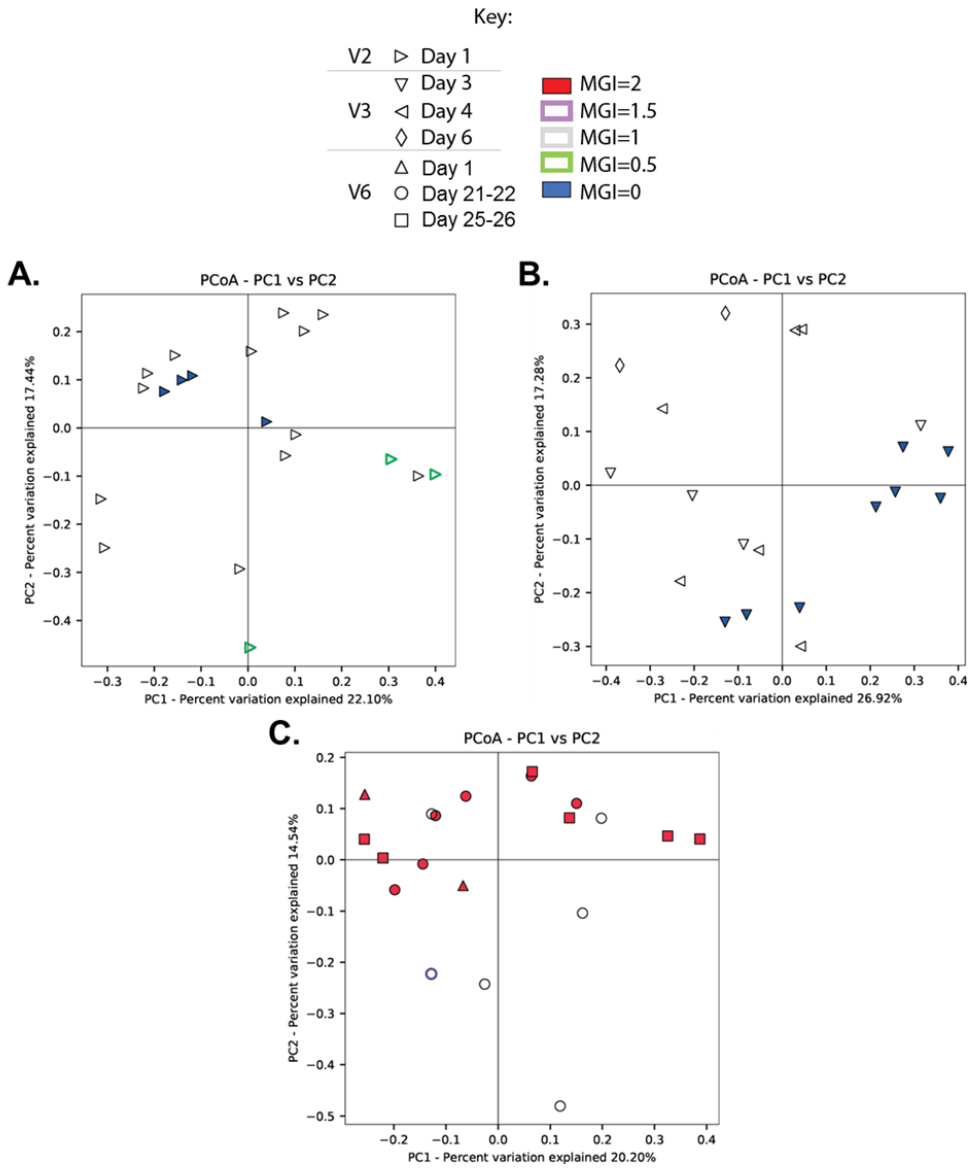


Figure 5.5 PCoA clustering of 16S rRNA sequencing samples by MGI score and visit.

16S rRNA sequencing reads were categorized into distinct operational taxonomical units (OTUs) by mapping to the HOMD (v14.51) Reference database using standard Qiime scripts. Beta-diversity between samples was measured through a Bray-Curtis dissimilarity analysis of the different OTUs within samples. Coloring indicates severity of gingivitis by MGI score while the shape of each point relates the day on which sample is collected for A) Visit 2 (MGI p=0.025); B) Visit 3 (MGI p=0.012, timepoint p=0.007); and C) Visit 6 (MGI p=0.08, timepoint p=0.122).

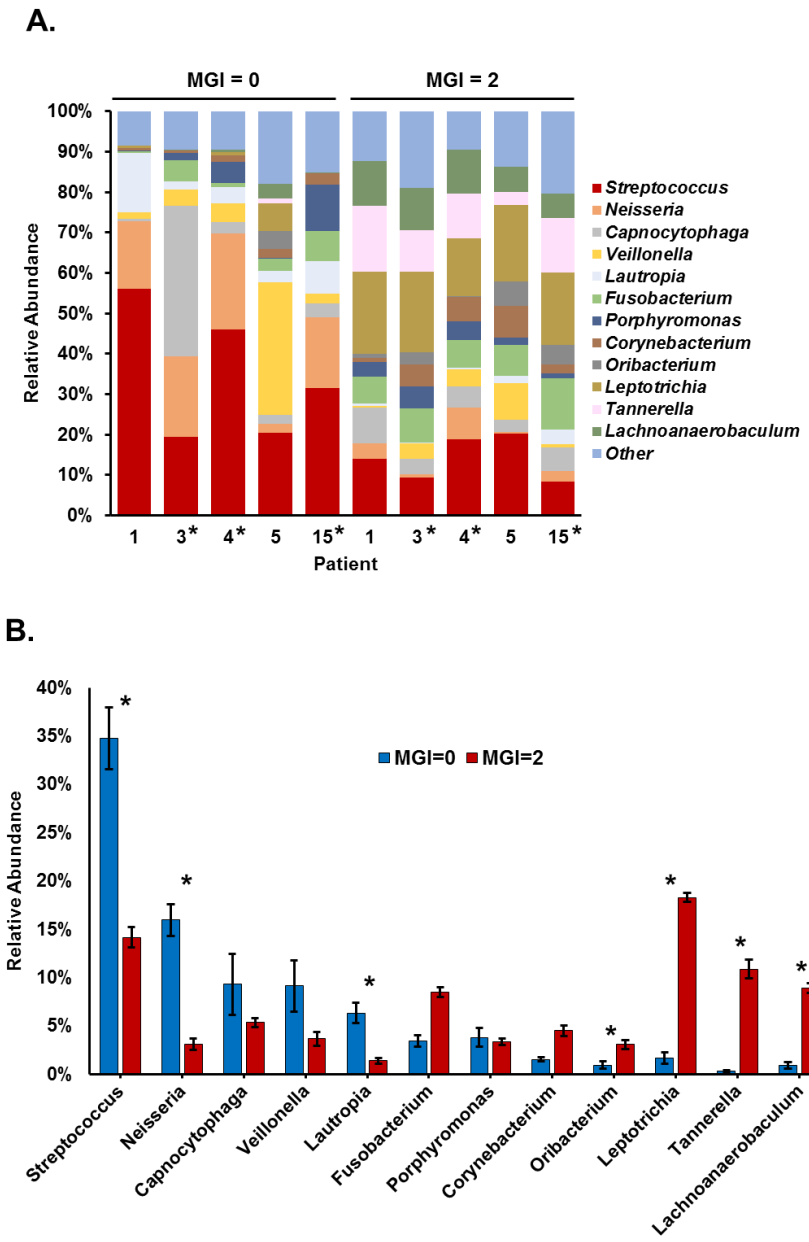
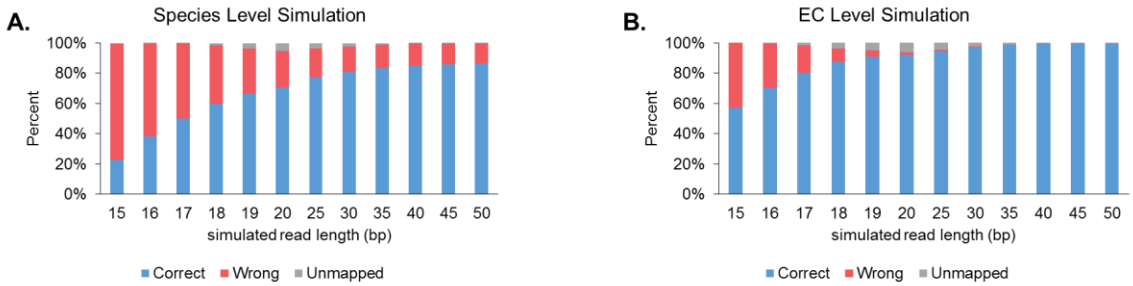


Figure 5.6 Composition of microbial communities from dental plaque samples assessed by 16S rRNA sequencing analysis.

A) Relative abundance of genera in the oral subginigival plaque community representing ≥ 1% within each subgingival plaque sample based on 16S rRNA sequencing. Samples further analyzed via RNAseq metatranscriptome analysis are indicated with an asterisk. B) Changes in average genus percent abundance across 5 patients (Patients 1, 3, 4, 5 and 15) from samples collected from teeth with an MGI score of 0 (clinical health) or an MGI score of 2 (clinical disease). Error bars represent standard error. Genera that change significantly from MGI=0 to MGI=2 samples are indicated with an asterisk ($p < 0.05$).



C.

Total read numbers:	Patient 3		Patient 4		Patient 15	
	MGI 0 (Visit 3)	MGI 2 (Visit 6)	MGI 0 (Visit 3)	MGI 2 (Visit 6)	MGI 0 (Visit 3)	MGI 2 (Visit 6)
Reads trimmed to ≥ 18bp	132,817,008	130,178,997	148,272,438	137,191,667	170,493,365	190,077,841
Reads mapping to HOMD	97,424,660 (73.35%)	101,708,693 (78.13%)	102,417,729 (69.07%)	97,693,664 (71.21%)	90,559,549 (53.12%)	108,188,692 (56.92%)
Reads mapping to hg19	13,729,648	5,906,255	27,454,933	20,966,727	16,626,553	21,085,323
Unmapped reads	26,932,467	25,111,769	28,384,583	22,134,200	70,567,177	68,787,668
HOMD Reads mapping to coding regions	40,437,690	42,118,406	42,547,518	31,200,298	36,043,037	36,727,265
Reads mapping to non-coding regions or ribosomal RNA	56,986,970	59,590,287	59,870,211	66,493,366	54,516,512	71,461,427
Average sequencing depth	66.9	29.7	74.0	31.8	45.3	58.2

Figure 5.7 Simulated mapping analysis and mapping statistics.

One million single end simulated reads were generated through wgsim from either A) the entire metatranscriptome; or B) a subset of genes with E.C. annotations. Mapping was performed using Bowtie2 with parameters used in the study to the entire transcriptome in both cases. Correct reads were identified by right mapping to the gene in which the simulated read originated. For species-level simulation, increasing the read length from the 30bp used in the analysis would only marginally increase reads for downstream analysis. For EC-level simulation, we achieve >80% correct mapping with read lengths of 18bp. C) Table showing read mapping statistics for each sample.

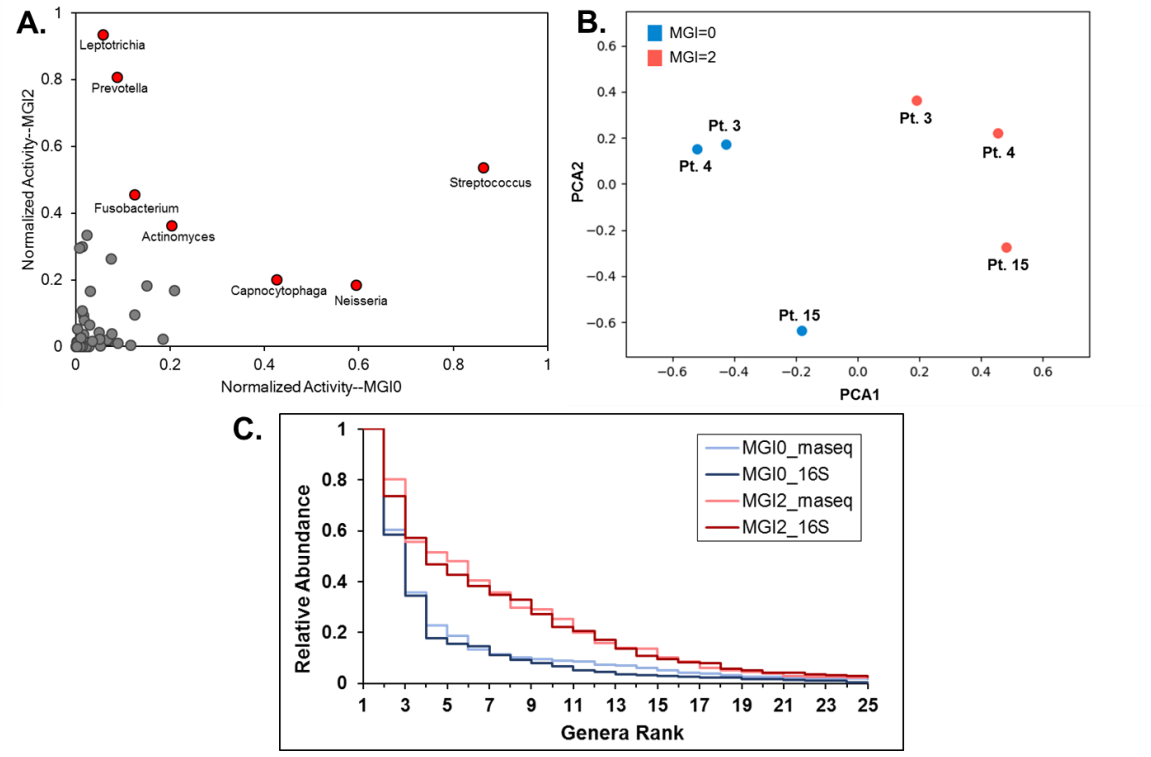


Figure 5.8 Comparison of transcriptomic data between disease states, patients, and taxonomic abundance.

A) Transcriptomic activity is normalized within-sample to the highest activity genus and aggregated by genus across the three samples for MGI0 samples (X-axis) and MGI2 samples (Y-axis), allowing a comparison of relative sample abundance between teeth at states of clinical health (MGI0) and clinical disease (MGI2). Red-colored data points indicate genera with the highest relative abundance in either MGI=0 or MGI=2 samples B) PCA analysis of RNAseq reads originating from each patient sample colored by MGI score. C) Rank abundance curves are shown for MGI=0 and MGI=2 samples for both taxonomic and transcriptomic analyses.

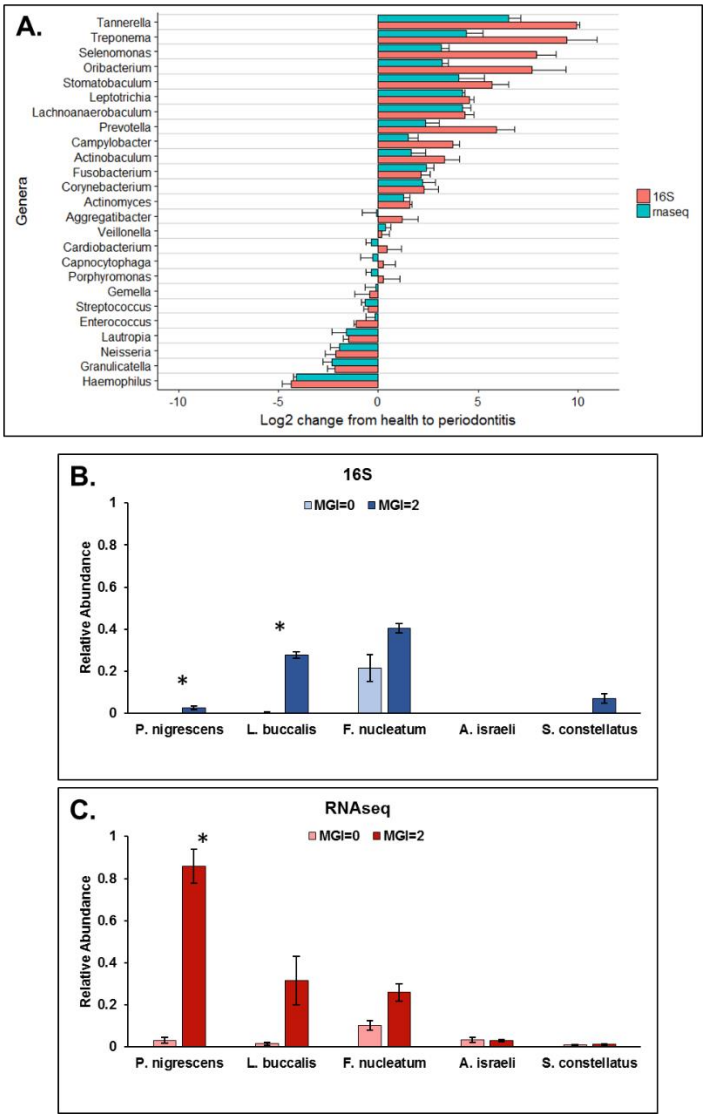


Figure 5.9 Genera and species level comparison for 16S rRNA and RNA sequencing data.

A.) Genera comparison using log2 fold change from patient-matched healthy and periodontitis sample in both 16S (red) and RNAseq analyses (blue). B) taxonomic abundances from MGI=0 and MGI=2 samples and C) transcriptomic abundances from MGI=0 and MGI=2 samples. (*) indicates significant difference ($p < 0.05$). Reads were mapped to genes and aggregated based upon species of origin. Relative abundance was calculated by dividing each species count by the highest count species within each sample and averaged across all three samples.

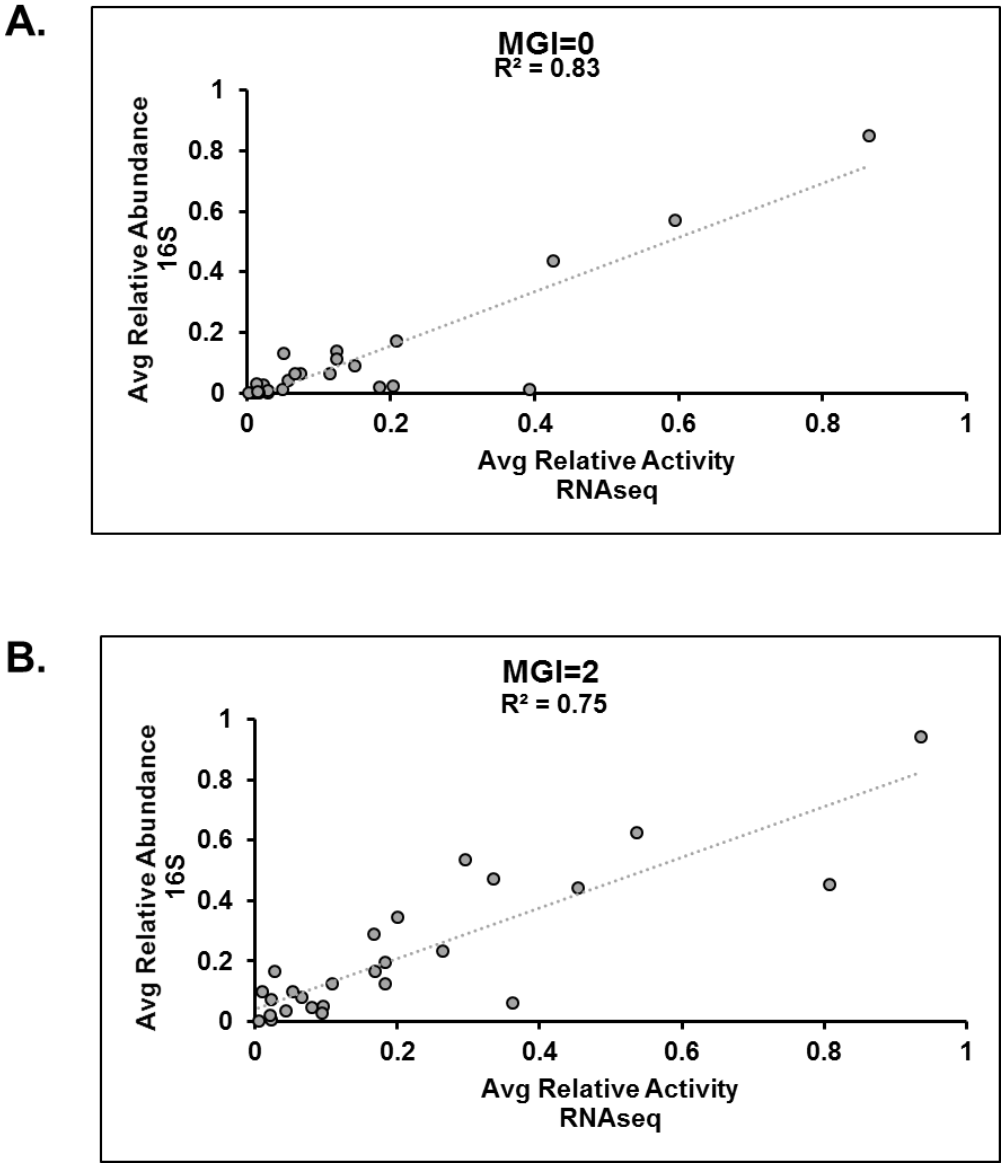


Figure 5.10 Genera abundance comparison between 16S rRNA and RNAseq data.

Shown in A) is the average relative abundance of the most abundant genera in MGI=0 samples are plotted for both 16S rRNA and RNAseq data; shown in B) is the average relative abundance of the most abundant genera in MGI=2 samples are plotted for both 16S rRNA and RNAseq data.

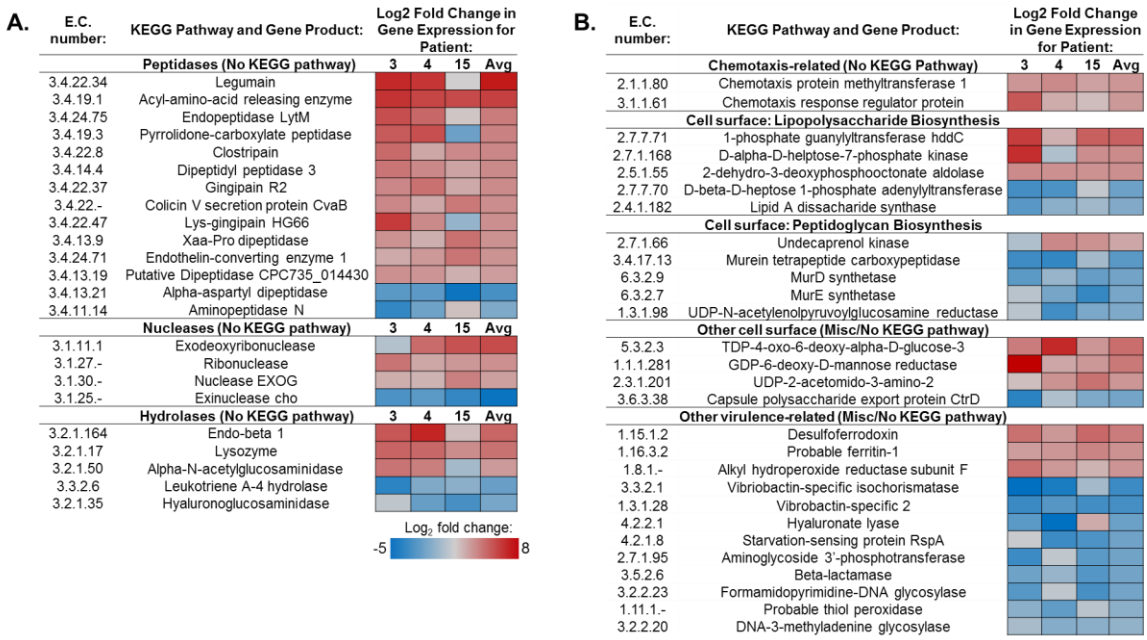


Figure 5.11 Virulence-related genes with significant differential expression between health and disease.

Differential expression between samples collected from clinically diseased (MGI=2, visit 6) and cliinically healthy (MGI=0, visit 3) teeth were compared, and reported as significant if $p \leq 0.05$. Log_2 fold change in gene expression was calculated for patients 3, 4 and 15 individually, or pooled across all three patients. Shown here are A) Hydrolytic enzymes or B) other virulence-related genes that are significantly up- or downregulated in samples from clinically diseased individuals relative those from healthy patients.

Organism:	Locus Tag:	Gene Product	Log_2 Fold change:	P-value:
Leptotrichia buccalis	lbuc_c_1_2002	DNA protection during starvation protein 1	8.4	0.000
	lbuc_c_1_1321	Response regulator MprA	8.3	0.000
	lbuc_c_1_356	Penicillin-binding protein 2	7.6	0.001
	lbuc_c_1_1097	Oligoendopeptidase F homolog	7.1	0.002
	lbuc_c_1_87	Toxin YoeB	6.6	0.005
	lbuc_c_1_416	Ferrous iron transport protein B	5.7	0.011
	lbuc_c_1_16	Multidrug export protein MepA	5.6	0.024
	lbuc_c_1_814	Extracellular serine protease	4.8	0.029
Prevotella nigrescens	pnig_c_9_1227	Multidrug resistance ABC transporter	9.1	0.000
	pnig_c_5_880	Multidrug export protein MepA	8.8	0.000
	pnig_c_4_764	Superkiller protein 3	8.8	0.000
	pnig_c_14_1461	Multidrug resistance protein NorM	6.6	0.005
	pnig_c_16_1584	Endothelin-converting enzyme 1	6.2	0.001
	pnig_c_13_1445	Thiol protease/hemagglutinin PrtT	5.9	0.003
	pnig_c_18_1640	Collagenase	5.6	0.008
	pnig_c_3_568	Ferrous iron transport protein B	5.5	0.005
	pnig_c_3_496	Protease PrtH	5.4	0.007
	pnig_c_28_1930	Xaa-Pro dipeptidase	5.3	0.007
	pnig_c_10_1306	Putative surface protein bspA-like	5.3	0.011
	pnig_c_3_540	Gingipain R1	4.4	0.040
Fusobacterium nucleatum	fnuc420_c_1_251	Cytosol non-specific dipeptidase	8.0	0.000
	fnuc2539_c_1_1214	Oligoendopeptidase F homolog	7.8	0.001
	fnuc2539_c_1_2112	Exodeoxyribonuclease	7.6	0.001
	fnuc420_c_1_145	Thermostable carboxypeptidase 2	7.5	0.001
	fnuc420_c_6_1159	Ferric transport protein FbpB	7.4	0.002
	fnuc2539_c_1_1239	Iron import ATP-binding/permease protein IrtA	7.0	0.004
	fnuc2539_c_1_1278	Probable multidrug resistance protein YoeA	6.9	0.004
	fnucp_c_3_1299	Penicillin-binding protein 1C	6.8	0.003
	fnuc2539_c_1_1432	Probable multidrug resistance protein NorM	6.7	0.006
	fnuc2539_c_1_924	Endoribonuclease YbeY	6.6	0.006
	fnuc2539_c_1_1998	Filamentous hemagglutinin	6.5	0.002
	fnuc2539_c_1_1429	Ferrous iron transport protein B	5.3	0.010
Streptococcus constellatus	scon_c_1_159	Alkyl hydroperoxide reductase subunit	5.8	0.018
	scon_c_3_1434	Regulatory protein spx	5.7	0.019
	scon_c_2_864	Uncharacterized protease YdeA	4.9	0.047
	scon_c_2_968	Ferrous iron transport protein B	4.5	0.042
Actinomyces israelii	aisr_c_21_2065	Virulence-associated protein I	6.4	0.008
	aisr_c_31_2454	Ribonuclease VapC35	5.0	0.042
	aisr_c_11_1426	Putative deoxyribonuclease RhsC	3.9	0.047

Table 5.2 Upregulated virulence-related genes of representative periodontal pathogens from the five most abundant genera during gingivitis (MGI=2).

Species analyzed include *L. buccalis, P. nigrescens, S. constellatus, F. nucleatum,* and *A. israelii.* Read counts were normalized for all ORFs within the metatranscriptome, and significantly differentially expressed genes pooled across three patients (3, 4 and 15) were determined by DESeq2. The Log_2 Fold changes in gene expression shown here represent differential expression between samples collected from teeth with gingivitis (MGI=2, visit 6) relative to healthy teeth (MGI=0, visit 3).

References

1. Moraes, F. & Góes, A. A decade of human genome project conclusion: Scientific diffusion about our genome knowledge. *Biochem Mol Biol Educ* **44**, 215–223 (2016).
2. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat. Methods* **5**, 16–18 (2008).
3. Galperin, M. Y. The Molecular Biology Database Collection: 2008 update. *Nucleic Acids Res.* **36**, D2-4 (2008).
4. Bouadjenek, M. R., Zobel, J. & Verspoor, K. Automated assessment of biological database assertions using the scientific literature. *BMC Bioinformatics* **20**, 216 (2019).
5. Rosenblatt, F. *PRINCIPLES OF NEURODYNAMICS. PERCEPTRONS AND THE THEORY OF BRAIN MECHANISMS*. https://apps.dtic.mil/docs/citations/AD0256582 (1961).
6. Plaut, D. C. & And Others. *Experiments on Learning by Back Propagation*. (1986).
7. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).
8. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
9. Addou, S., Rentzsch, R., Lee, D. & Orengo, C. A. Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. *J. Mol. Biol.* **387**, 416–430 (2009).
10. Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463–467 (2011).
11. Jaakkola, T., Diekhans, M. & Haussler, D. A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.* **7**, 95–114 (2000).
12. Leslie, C. S., Eskin, E., Cohen, A., Weston, J. & Noble, W. S. Mismatch string kernels for discriminative protein classification. *Bioinformatics* **20**, 467–476 (2004).
13. Qiu, J., Hue, M., Ben-Hur, A., Vert, J.-P. & Noble, W. S. A structural alignment kernel for protein structures. *Bioinformatics* **23**, 1090–1098 (2007).
14. Rifaioglu, A. S. *et al.* Large-scale automated function prediction of protein sequences and an experimental case study validation on PTEN transcript variants. *Proteins* **86**, 135–151 (2018).
15. Cao, R. *et al.* ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. *Molecules* **22**, (2017).
16. Szalkai, B. & Grolmusz, V. Near perfect protein multi-label classification with deep neural networks. *Methods* **132**, 50–56 (2018).
17. Kulmanov, M., Khan, M. A., Hoehndorf, R. & Wren, J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34**, 660–668 (2018).
18. Bileschi, M. L. *et al.* Using Deep Learning to Annotate the Protein Universe. *bioRxiv* 626507 (2019) doi:10.1101/626507.

19. Sureyya Rifaioglu, A., Doğan, T., Jesus Martin, M., Cetin-Atalay, R. & Atalay, V. DEEPred: Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks. *Sci Rep* **9**, 7344 (2019).

20. Dalkiran, A. *et al.* ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics* **19**, 334 (2018).

21. Strodthoff, N., Wagner, P., Wenzel, M. & Samek, W. Universal Deep Sequence Models for Protein Classification. *bioRxiv* 704874 (2019) doi:10.1101/704874.

22. Wan, C. & Jones, D. T. Improving protein function prediction with synthetic feature samples created by generative adversarial networks. *bioRxiv* 730143 (2019) doi:10.1101/730143.

23. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).

24. Fox, R. J. *et al.* Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **25**, 338–344 (2007).

25. Li, Y. *et al.* A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.* **25**, 1051–1056 (2007).

26. Liao, J. *et al.* Engineering proteinase K using machine learning and synthetic genes. *BMC Biotechnol.* **7**, 16 (2007).

27. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E193-201 (2013).

28. Bedbrook, C. N. *et al.* Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat. Methods* **16**, 1176–1184 (2019).

29. Saito, Y. *et al.* Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins. *ACS Synth Biol* **7**, 2014–2022 (2018).

30. Biswas, S. *et al.* Toward machine-guided design of proteins. *bioRxiv* 337154 (2018) doi:10.1101/337154.

31. Khurana, S. *et al.* DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* **34**, 2605–2613 (2018).

32. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* **32**, D120-121 (2004).

33. Kwasigroch, J. M., Gilis, D., Dehouck, Y. & Rooman, M. PoPMuSiC, rationally designing point mutations in protein structures. *Bioinformatics* **18**, 1701–1702 (2002).

34. Dehouck, Y. *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* **25**, 2537–2543 (2009).

35. Giollo, M., Martin, A. J. M., Walsh, I., Ferrari, C. & Tosatto, S. C. E. NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. *BMC Genomics* **15 Suppl 4**, S7 (2014).

36. Ng, P. dna2vec: Consistent vector representations of variable-length k-mers. *arXiv:1701.06279 [cs, q-bio, stat]* (2017).

37. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]* (2013).

38. Yang, K. K., Wu, Z., Bedbrook, C. N. & Arnold, F. H. Learned protein embeddings for machine learning. *Bioinformatics* **34**, 4138 (2018).

39. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* (2019) doi:10.1038/s41592-019-0598-1.

40. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv* 622803 (2019) doi:10.1101/622803.

41. Heinzinger, M. *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* **20**, 723 (2019).

42. Li, Z., Yang, Y., Faraggi, E., Zhan, J. & Zhou, Y. Direct prediction of profiles of sequences compatible to a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles. *Proteins* **82**, 2565–2573 (2014).

43. O'Connell, J. *et al.* SPIN2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins* **86**, 629–633 (2018).

44. Torng, W. & Altman, R. B. 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinformatics* **18**, 302 (2017).

45. Boomsma, W. & Frellsen, J. Spherical convolutions and their application in molecular modelling. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 3433–3443 (Curran Associates, Inc., 2017).

46. Weiler, M., Geiger, M., Welling, M., Boomsma, W. & Cohen, T. S. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. in *Advances in Neural Information Processing Systems 31* (eds. Bengio, S. et al.) 10381–10392 (Curran Associates, Inc., 2018).

47. Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 16856–16865 (2019).

48. AlQuraishi, M. End-to-End Differentiable Learning of Protein Structure. *Cell Syst* **8**, 292-301.e3 (2019).

49. Senior, A. W. *et al.* Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins* **87**, 1141–1148 (2019).

50. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **35**, 4862–4865 (2019).

51. Sinai, S., Kelsic, E., Church, G. M. & Nowak, M. A. Variational auto-encoding of protein sequences. *arXiv:1712.03346 [cs, q-bio]* (2018).

52. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).

53. Costello, Z. & Martin, H. G. How to Hallucinate Functional Proteins. *arXiv:1903.00458 [q-bio]* (2019).

54. Davidsen, K. *et al.* Deep generative models for T cell receptor protein sequences. *Elife* **8**, (2019).

55. Greener, J. G., Moffat, L. & Jones, D. T. Design of metalloproteins and novel protein folds using variational autoencoders. *Sci Rep* **8**, 16189 (2018).

56. Riesselman, A. *et al.* Accelerating Protein Design Using Autoregressive Generative Models. *bioRxiv* 757252 (2019) doi:10.1101/757252.

57. Goodfellow, I. *et al.* Generative Adversarial Nets. in *Advances in Neural Information Processing Systems 27* (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q.) 2672–2680 (Curran Associates, Inc., 2014).

58. Gupta, A. & Zou, J. Feedback GAN (FBGAN) for DNA: a Novel Feedback-Loop Architecture for Optimizing Protein Functions. *arXiv:1804.01694 [cs, q-bio]* (2018).

59. Karimi, M., Zhu, S., Cao, Y. & Shen, Y. De Novo Protein Design for Novel Folds using Guided Conditional Wasserstein Generative Adversarial Networks (gcWGAN). *bioRxiv* 769919 (2019) doi:10.1101/769919.

60. Jensen, P. F. *et al.* Structure and Dynamics of a Promiscuous Xanthan Lyase from Paenibacillus nanensis and the Design of Variants with Increased Stability and Activity. *Cell Chem Biol* **26**, 191-202.e6 (2019).

61. Windle, C. L. *et al.* Extending enzyme molecular recognition with an expanded amino acid alphabet. *Proc Natl Acad Sci U S A* **114**, 2610–2615 (2017).

62. Studer, S. *et al.* Evolution of a highly active and enantiospecific metalloenzyme from short peptides. *Science* **362**, 1285–1288 (2018).

63. Trudeau, D. L., Kaltenbach, M. & Tawfik, D. S. On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. *Mol. Biol. Evol.* **33**, 2633–2641 (2016).

64. Potapov, V., Cohen, M. & Schreiber, G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng. Des. Sel.* **22**, 553–560 (2009).

65. Borgo, B. & Havranek, J. J. Automated selection of stabilizing mutations in designed and natural proteins. *Proc Natl Acad Sci U S A* **109**, 1494–1499 (2012).

66. Steinbrecher, T. & Elstner, M. QM and QM/MM simulations of proteins. *Methods Mol. Biol.* **924**, 91–124 (2013).

67. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).

68. Jia, L., Yarlagadda, R. & Reed, C. C. Structure Based Thermostability Prediction Models for Protein Single Point Mutations with Machine Learning Tools. *PLoS ONE* **10**, e0138022 (2015).

69. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8852–8858 (2019).

70. Jiménez, J., Škalič, M., Martínez-Rosell, G. & De Fabritiis, G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J Chem Inf Model* **58**, 287–296 (2018).

71. Amidi, A. *et al.* EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation. *PeerJ* **6**, e4750 (2018).

72. Wang, J., Cao, H., Zhang, J. Z. H. & Qi, Y. Computational Protein Design with Deep Learning Neural Networks. *Sci Rep* **8**, (2018).

73. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst* **6**, 116-124.e3 (2018).

74. Costantini, L. M., Subach, O. M., Jaureguiberry-bravo, M., Verkhusha, V. V. & Snapp, E. L. Cysteineless non-glycosylated monomeric blue fluorescent protein, secBFP2, for studies in the eukaryotic secretory pathway. *Biochem. Biophys. Res. Commun.* **430**, 1114–1119 (2013).

75. Kather, I., Jakob, R. P., Dobbek, H. & Schmid, F. X. Increased folding stability of TEM-1 beta-lactamase by in vitro selection. *J. Mol. Biol.* **383**, 238–251 (2008).

76. Zimmerman, M. I. *et al.* Prediction of New Stabilizing Mutations Based on Mechanistic Insights from Markov State Models. *ACS Cent Sci* **3**, 1311–1321 (2017).

77. Farzaneh, S. *et al.* Implication of Ile-69 and Thr-182 residues in kinetic characteristics of IRT-3 (TEM-32) beta-lactamase. *Antimicrob. Agents Chemother.* **40**, 2434–2436 (1996).

78. Orencia, M. C., Yoon, J. S., Ness, J. E., Stemmer, W. P. & Stevens, R. C. Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nat. Struct. Biol.* **8**, 238–242 (2001).

79. Joosten, R. P. *et al.* PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J Appl Crystallogr* **42**, 376–384 (2009).

80. Dolinsky, T. J. *et al.* PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35**, W522-525 (2007).

81. Mitternacht, S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res* **5**, 189 (2016).

82. Jacquier, H. *et al.* Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13067–13072 (2013).

83. Cabantous, S., Rogers, Y., Terwilliger, T. C. & Waldo, G. S. New molecular reporters for rapid protein folding assays. *PLoS ONE* **3**, e2387 (2008).

84. Bratulic, S., Gerber, F. & Wagner, A. Mistranslation drives the evolution of robustness in TEM-1 β-lactamase. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 12758–12763 (2015).

85. Guthrie, V. B., Allen, J., Camps, M. & Karchin, R. Network models of TEM β-lactamase mutations coevolving under antibiotic selection show modular structure and anticipate evolutionary trajectories. *PLoS Comput. Biol.* **7**, e1002184 (2011).

86. Mori, Y., Kanda, H. & Notomi, T. Loop-mediated isothermal amplification (LAMP): recent progress in research and development. *Journal of infection and chemotherapy* **19**, 404–411 (2013).

87. Notomi, T. *et al.* Loop-mediated isothermal amplification of DNA. *Nucleic acids research* **28**, e63–e63 (2000).

88. Lizardi, P. M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nature genetics* **19**, 225 (1998).

89. Zhang, D. Y., Zhang, W., Li, X. & Konomi, Y. Detection of rare DNA targets by isothermal ramification amplification. *Gene* **274**, 209–216 (2001).

90. Lawyer, F. C. *et al.* High-level expression, purification, and enzymatic characterization of full-length Thermus aquaticus DNA polymerase and a truncated form deficient in 5' to 3' exonuclease activity. *PCR Methods Appl.* **2**, 275–287 (1993).

91. Craw, P. & Balachandran, W. Isothermal nucleic acid amplification technologies for point-of-care diagnostics: a critical review. *Lab on a Chip* **12**, 2469–2486 (2012).

92. R. Hartman, M. *et al.* Point-of-care nucleic acid detection using nanotechnology. *Nanoscale* **5**, 10141–10154 (2013).

93. Gill, P. & Ghaemi, A. Nucleic acid isothermal amplification technologies—a review. *Nucleosides, Nucleotides and Nucleic Acids* **27**, 224–243 (2008).

94. Njiru, Z. K. Loop-mediated isothermal amplification technology: towards point of care diagnostics. *PLoS Negl Trop Dis* **6**, e1572 (2012).

95. Zhao, Y., Chen, F., Li, Q., Wang, L. & Fan, C. Isothermal amplification of nucleic acids. *Chemical reviews* **115**, 12491–12545 (2015).

96. Asiello, P. J. & Baeumner, A. J. Miniaturized isothermal nucleic acid amplification, a review. *Lab Chip* **11**, 1420–1430 (2011).

97. Du, Y. *et al.* A Sweet Spot for Molecular Diagnostics: Coupling Isothermal Amplification and Strand Exchange Circuits to Glucometers. *Sci Rep* **5**, 11039 (2015).

98. Jiang, Y. S. *et al.* Robust strand exchange reactions for the sequence-specific, real-time detection of nucleic acid amplicons. *Anal. Chem.* **87**, 3314–3320 (2015).

99. Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome research* **11**, 1095–1099 (2001).

100. Jiang, Y., Li, B., Milligan, J. N., Bhadra, S. & Ellington, A. D. Real-time detection of isothermal amplification reactions with thermostable catalytic hairpin assembly. *Journal of the American Chemical Society* **135**, 7430–7433 (2013).

101. Bhadra, S. *et al.* Real-time sequence-validated loop-mediated isothermal amplification assays for detection of Middle East respiratory syndrome coronavirus (MERS-CoV). *PLoS ONE* **10**, e0123126 (2015).

102. Kiefer, J. R. *et al.* Crystal structure of a thermostable Bacillus DNA polymerase l large fragment at 2.1 \AA resolution. *Structure* **5**, 95–108 (1997).

103. Blanco, L. & Salas, M. Characterization and purification of a phage phi 29-encoded DNA polymerase required for the initiation of replication. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5325–5329 (1984).

104. Blanco, L. *et al.* Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **264**, 8935–8940 (1989).

105. McClary, J., Ye, S. Y., Hong, G. F. & Witney, F. Sequencing with the large fragment of DNA polymerase I from Bacillus stearothermophilus. *DNA Seq.* **1**, 173–180 (1991).

106. Ye, S. Y. & Hong, G. F. Heat-stable DNA polymerase I large fragment resolves hairpin structure in DNA sequencing. *Sci Sin B* **30**, 503–506 (1987).

107. Zhang, K. *et al.* Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **24**, 680–686 (2006).

108. Hsieh, K., Mage, P. L., Csordas, A. T., Eisenstein, M. & Soh, H. T. Simultaneous elimination of carryover contamination and detection of DNA with uracil-DNA-glycosylase-supplemented loop-mediated isothermal amplification (UDG-LAMP). *Chem. Commun. (Camb.)* **50**, 3747–3749 (2014).

109. Njiru, Z. K. *et al.* African trypanosomiasis: sensitive and rapid detection of the sub-genus Trypanozoon by loop-mediated isothermal amplification (LAMP) of parasite DNA. *Int. J. Parasitol.* **38**, 589–599 (2008).

110. Suzuki, R. *et al.* Heat denaturation increases the sensitivity of the cytomegalovirus loop-mediated isothermal amplification method. *Microbiol. Immunol.* **54**, 466–470 (2010).

111. Verkooyen, R. P. *et al.* Detection of PCR inhibitors in cervical specimens by using the AMPLICOR Chlamydia trachomatis assay. *J. Clin. Microbiol.* **34**, 3072–3074 (1996).

112. Modak, S. S. *et al.* Rapid Point-of-Care Isothermal Amplification Assay for the Detection of Malaria without Nucleic Acid Purification. *Infect Dis (Auckl)* **9**, 1–9 (2016).

113. Fereidouni, S. R. *et al.* Sample preparation for avian and porcine influenza virus cDNA amplification simplified: Boiling vs. conventional RNA extraction. *J. Virol. Methods* **221**, 62–67 (2015).

114. Queipo-Ortuño, M. I., De Dios Colmenero, J., Macias, M., Bravo, M. J. & Morata, P. Preparation of bacterial DNA template by boiling and effect of immunoglobulin G as an inhibitor in real-time PCR for serum samples from patients with brucellosis. *Clin. Vaccine Immunol.* **15**, 293–296 (2008).

115. Chander, Y. *et al.* A novel thermostable polymerase for RNA and DNA loop-mediated isothermal amplification (LAMP). *Front Microbiol* **5**, 395 (2014).

116. Ignatov, K. B. *et al.* A strong strand displacement activity of thermostable DNA polymerase markedly improves the results of DNA amplification. *BioTechniques* **57**, 81–87 (2014).

117. Ghadessy, F. J., Ong, J. L. & Holliger, P. Directed evolution of polymerase function by compartmentalized self-replication. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4552–4557 (2001).

118. Baar, C. *et al.* Molecular breeding of polymerases for resistance to environmental inhibitors. *Nucleic Acids Res.* **39**, e51 (2011).

119. Chen, T. & Romesberg, F. E. Directed polymerase evolution. *FEBS Lett.* **588**, 219–229 (2014).

120. Meyer, A. J., Ellefson, J. W. & Ellington, A. D. Directed Evolution of a Panel of Orthogonal T7 RNA Polymerase Variants for in Vivo or in Vitro Synthetic Circuitry. *ACS Synth Biol* **4**, 1070–1076 (2015).

121. Ellefson, J. W. *et al.* Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science* **352**, 1590–1593 (2016).

122. Povilaitis, T., Alzbutas, G., Sukackaite, R., Siurkus, J. & Skirgaila, R. In vitro evolution of phi29 DNA polymerase using isothermal compartmentalized self replication technique. *Protein Eng. Des. Sel.* **29**, 617–628 (2016).

123. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).

124. Temin, H. M. & Mizutani, S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* **226**, 1211–1213 (1970).

125. Baltimore, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* **226**, 1209–1211 (1970).

126. Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**, 3353–3362 (1990).

127. Boeke, J. D. & Stoye, J. P. Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements. in *Retroviruses* (eds. Coffin, J. M., Hughes, S. H. & Varmus, H. E.) (Cold Spring Harbor Laboratory Press, 1997).

128. Darnell, J. E. & Doolittle, W. F. Speculations on the early course of evolution. *Proc Natl Acad Sci U S A* **83**, 1271–1275 (1986).

129. Nakamura, T. M. *et al.* Telomerase Catalytic Subunit Homologs from Fission Yeast and Human. *Science* **277**, 955–959 (1997).

130. Meyerhans, A. *et al.* Temporal fluctuations in HIV quasispecies in vivo are not reflected by sequential HIV isolations. *Cell* **58**, 901–910 (1989).

131. Kunkel, T. A. & Bebenek, K. DNA replication fidelity. *Annu. Rev. Biochem.* **69**, 497–529 (2000).

132. Chen, T. *et al.* Evolution of thermophilic DNA polymerases for the recognition and amplification of C2'-modified DNA. *Nat Chem* **8**, 556–562 (2016).

133. Schultz, H. J. *et al.* Taq DNA Polymerase Mutants and 2′-Modified Sugar Recognition. *Biochemistry* **54**, 5999–6008 (2015).

134. Burmeister, P. E. *et al.* Direct in vitro selection of a 2'-O-methyl aptamer to VEGF. *Chem. Biol.* **12**, 25–33 (2005).

135. Ghadessy, F. J. *et al.* Generic expansion of the substrate spectrum of a DNA polymerase by directed evolution. *Nat. Biotechnol.* **22**, 755–759 (2004).

136. Sagner, G., Rüger, R. & Kessler, C. Rapid filter assay for the detection of DNA polymerase activity: direct identification of the gene for the DNA polymerase from Thermus aquaticus. *Gene* **97**, 119–123 (1991).

137. Korolev, S., Nayal, M., Barnes, W. M., Di Cera, E. & Waksman, G. Crystal structure of the large fragment of Thermus aquaticus DNA polymerase I at 2.5-A resolution: structural basis for thermostability. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 9264–9268 (1995).

138. Jiang, Y. (Sherry), Li, B., Milligan, J. N., Bhadra, S. & Ellington, A. D. Real-time detection of isothermal amplification reactions with thermostable catalytic hairpin assembly. *J Am Chem Soc* **135**, 7430–7433 (2013).

139. Nelson, J. R. *et al.* TempliPhi, phi29 DNA polymerase based rolling circle amplification of templates for DNA sequencing. *BioTechniques* **Suppl**, 44–47 (2002).

140. Reagin, M. J. *et al.* TempliPhi: A sequencing template preparation procedure that eliminates overnight cultures and DNA purification. *J Biomol Tech* **14**, 143–148 (2003).

141. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**, 14508–14513 (2012).

142. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).

143. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).

144. Ramalingam, N., San, T. C., Kai, T. J., Mak, M. Y. M. & Gong, H.-Q. Microfluidic devices harboring unsealed reactors for real-time isothermal helicase-dependent amplification. *Microfluid Nanofluid* **7**, 325 (2009).

145. Mahalanabis, M., Do, J., ALMuayad, H., Zhang, J. Y. & Klapperich, C. M. An integrated disposable device for DNA extraction and helicase dependent amplification. *Biomed Microdevices* **12**, 353–359 (2010).

146. Borysiak, M. D., Kimura, K. W. & Posner, J. D. NAIL: Nucleic Acid detection using Isotachophoresis and Loop-mediated isothermal amplification. *Lab Chip* **15**, 1697–1707 (2015).

147. Kim, T.-H., Park, J., Kim, C.-J. & Cho, Y.-K. Fully integrated lab-on-a-disc for nucleic acid analysis of food-borne pathogens. *Anal. Chem.* **86**, 3841–3848 (2014).

148. Wu, Q. *et al.* Integrated glass microdevice for nucleic acid purification, loop-mediated isothermal amplification, and online detection. *Anal. Chem.* **83**, 3336–3342 (2011).

149. Marincevic-Zuniga, Y., Gustavsson, I. & Gyllensten, U. Multiply-primed rolling circle amplification of human papillomavirus using sequence-specific primers. *Virology* **432**, 57–62 (2012).

150. Kiefer, J. R., Mao, C., Braman, J. C. & Beese, L. S. Visualizing DNA replication in a catalytically active Bacillus DNA polymerase crystal. *Nature* **391**, 304–307 (1998).

151. Li, Y., Korolev, S. & Waksman, G. Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of Thermus aquaticus DNA polymerase I: structural basis for nucleotide incorporation. *EMBO J.* **17**, 7514–7525 (1998).

152. Huff, J. W., Sastry, K. S., Gordon, M. P. & Wacker, W. E. THE ACTION OF METAL IONS ON TOBACCO MOSAIC VIRUS RIBONUCLEIC ACID. *Biochemistry* **3**, 501–506 (1964).

153. Malyshev, D. A. *et al.* Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. *Proc Natl Acad Sci U S A* **109**, 12005–12010 (2012).

154. Yang, Z., Chen, F., Alvarado, J. B. & Benner, S. A. Amplification, Mutation, and Sequencing of a Six-Letter Synthetic Genetic System. *J. Am. Chem. Soc.* **133**, 15105–15112 (2011).

155. Yamashige, R. *et al.* Highly specific unnatural base pair systems as a third base pair for PCR amplification. *Nucleic Acids Res.* **40**, 2793–2806 (2012).

156. Schöning, K. *et al.* Chemical etiology of nucleic acid structure: the alpha-threofuranosyl-(3'-->2') oligonucleotide system. *Science* **290**, 1347–1351 (2000).

157. Pinheiro, V. B. *et al.* Synthetic genetic polymers capable of heredity and evolution. *Science* **336**, 341–344 (2012).

158. McDaniel, J. R. *et al.* Identification of tumor-reactive B cells and systemic IgG in breast cancer based on clonal frequency in the sentinel lymph node. *Cancer Immunol. Immunother.* **67**, 729–738 (2018).

159. Bhadra, S. *et al.* Cellular reagents for diagnostics and synthetic biology. *PLOS ONE* **13**, e0201681 (2018).

160. Knight, R. D., Freeland, S. J. & Landweber, L. F. Rewiring the keyboard: evolvability of the genetic code. *Nat. Rev. Genet.* **2**, 49–58 (2001).

161. Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56**, 229–264 (1992).

162. Schultz, D. W. & Yarus, M. Transfer RNA mutation and the malleability of the genetic code. *J. Mol. Biol.* **235**, 1377–1380 (1994).

163. Andersson, S. G. & Kurland, C. G. Reductive evolution of resident genomes. *Trends Microbiol.* **6**, 263–268 (1998).

164. Wong, J. T. Membership mutation of the genetic code: loss of fitness by tryptophan. *Proc. Natl. Acad. Sci. U.S.A.* **80**, 6303–6306 (1983).

165. Hoesl, M. G. *et al.* Chemical Evolution of a Bacterial Proteome. *Angew. Chem. Int. Ed. Engl.* **54**, 10030–10034 (2015).

166. Hammerling, M. J. *et al.* Bacteriophages use an expanded genetic code on evolutionary paths to higher fitness. *Nat. Chem. Biol.* **10**, 178–180 (2014).

167. Wang, L. & Schultz, P. G. A general approach for the generation of orthogonal tRNAs. *Chem. Biol.* **8**, 883–890 (2001).

168. Maranhao, A. C. & Ellington, A. D. Evolving Orthogonal Suppressor tRNAs To Incorporate Modified Amino Acids. *ACS Synth. Biol.* **6**, 108–119 (2017).

169. Sakamoto, K. *et al.* Genetic encoding of 3-iodo-L-tyrosine in Escherichia coli for single-wavelength anomalous dispersion phasing in protein crystallography. *Structure* **17**, 335–344 (2009).

170. Wang, Q. *et al.* Response and adaptation of Escherichia coli to suppression of the amber stop codon. *Chembiochem* **15**, 1744–1749 (2014).

171. Mukai, T. *et al.* Codon reassignment in the Escherichia coli genetic code. *Nucleic Acids Res.* **38**, 8188–8195 (2010).

172. Bacher, J. M., Bull, J. J. & Ellington, A. D. Evolution of phage with chemically ambiguous proteomes. *BMC Evol. Biol.* **3**, 24 (2003).

173. Isaacs, F. J. *et al.* Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. *Science* **333**, 348–353 (2011).

174. Hammerling, M. J. *et al.* Expanded Genetic Codes Create New Mutational Routes to Rifampicin Resistance in Escherichia coli. *Mol. Biol. Evol.* **33**, 2054–2063 (2016).

175. Thyer, R., Robotham, S. A., Brodbelt, J. S. & Ellington, A. D. Evolving tRNA(Sec) for efficient canonical incorporation of selenocysteine. *J. Am. Chem. Soc.* **137**, 46–49 (2015).

176. Lajoie, M. J. *et al.* Genomically recoded organisms expand biological functions. *Science* **342**, 357–360 (2013).

177. Blount, Z. D., Barrick, J. E., Davidson, C. J. & Lenski, R. E. Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* **489**, 513–518 (2012).

178. Quandt, E. M. *et al.* Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the Lenski evolution experiment. *eLife* **4**,.

179. Majiduddin, F. K. & Palzkill, T. Amino acid sequence requirements at residues 69 and 238 for the SME-1 beta-lactamase to confer resistance to beta-lactam antibiotics. *Antimicrob. Agents Chemother.* **47**, 1062–1067 (2003).

180. Gutierrez, A. *et al.* β-Lactam antibiotics promote bacterial mutagenesis via an RpoS-mediated reduction in replication fidelity. *Nat Commun* **4**, 1610 (2013).

181. Layton, J. C. & Foster, P. L. Error-prone DNA polymerase IV is regulated by the heat shock chaperone GroE in Escherichia coli. *J. Bacteriol.* **187**, 449–457 (2005).

182. Ishii, T. M., Kotlova, N., Tapsoba, F. & Steinberg, S. V. The long D-stem of the selenocysteine tRNA provides resilience at the expense of maximal function. *J. Biol. Chem.* **288**, 13337–13344 (2013).

183. Nilsson, M. & Rydén-Aulin, M. Glutamine is incorporated at the nonsense codons UAG and UAA in a suppressor-free Escherichia coli strain. *Biochim. Biophys. Acta* **1627**, 1–6 (2003).

184. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898 (2009).

185. Baggett, N. E., Zhang, Y. & Gross, C. A. Global analysis of translation termination in E. coli. *PLoS Genet.* **13**, e1006676 (2017).

186. Mora, L., Heurgué-Hamard, V., de Zamaroczy, M., Kervestin, S. & Buckingham, R. H. Methylation of bacterial release factors RF1 and RF2 is required for normal translation termination in vivo. *J. Biol. Chem.* **282**, 35638–35645 (2007).

187. Pagès, J.-M., James, C. E. & Winterhalter, M. The porin and the permeating antibiotic: a selective diffusion barrier in Gram-negative bacteria. *Nat. Rev. Microbiol.* **6**, 893–903 (2008).

188. Sun, S., Selmer, M. & Andersson, D. I. Resistance to β-lactam antibiotics conferred by point mutations in penicillin-binding proteins PBP3, PBP4 and PBP6 in Salmonella enterica. *PLoS ONE* **9**, e97202 (2014).

189. Kullik, I., Toledano, M. B., Tartaglia, L. A. & Storz, G. Mutational analysis of the redox-sensitive transcriptional regulator OxyR: regions important for oxidation and transcriptional activation. *J. Bacteriol.* **177**, 1275–1284 (1995).

190. Lacourciere, G. M., Levine, R. L. & Stadtman, T. C. Direct detection of potential selenium delivery proteins by using an Escherichia coli strain unable to incorporate selenium from selenite into proteins. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 9150–9153 (2002).

191. Wannier, T. M. *et al.* Adaptive evolution of genomically recoded Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3090–3095 (2018).

192. Batchelor, E., Walthers, D., Kenney, L. J. & Goulian, M. The Escherichia coli CpxA-CpxR envelope stress response system regulates expression of the porins ompF and ompC. *J. Bacteriol.* **187**, 5723–5731 (2005).

193. Schwartz, C. J. *et al.* IscR, an Fe-S cluster-containing transcription factor, represses expression of Escherichia coli genes encoding Fe-S cluster assembly proteins. *Proc Natl Acad Sci U S A* **98**, 14895–14900 (2001).

194. Chang, D.-E., Shin, S., Rhee, J.-S. & Pan, J.-G. Acetate Metabolism in a pta Mutant of Escherichia coli W3110: Importance of Maintaining Acetyl Coenzyme A Flux for Growth and Survival. *J Bacteriol* **181**, 6656–6663 (1999).

195. Sezonov, G., Joseleau-Petit, D. & D'Ari, R. Escherichia coli Physiology in Luria-Bertani Broth. *J Bacteriol* **189**, 8746–8749 (2007).

196. Kuznetsov, G. *et al.* Optimizing complex phenotypes through model-guided multiplex genome engineering. *Genome Biol.* **18**, 100 (2017).

197. Yang, Y. L. & Polisky, B. Suppression of ColE1 high-copy-number mutants by mutations in the polA gene of Escherichia coli. *J Bacteriol* **175**, 428–437 (1993).

198. Tack, D. S. *et al.* Addicting diverse bacteria to a noncanonical amino acid. *Nat. Chem. Biol.* **12**, 138–140 (2016).

199. Stojanoski, V. *et al.* A triple mutant in the Ω-loop of TEM-1 β-lactamase changes the substrate profile via a large conformational change and an altered general base for catalysis. *J. Biol. Chem.* **290**, 10382–10394 (2015).

200. Palzkill, T., Le, Q. Q., Venkatachalam, K. V., LaRocco, M. & Ocera, H. Evolution of antibiotic resistance: several different amino acid substitutions in an active site loop alter the substrate profile of beta-lactamase. *Mol. Microbiol.* **12**, 217–229 (1994).

201. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453–1462 (1997).

202. Salverda, M. L. M., De Visser, J. A. G. M. & Barlow, M. Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiol. Rev.* **34**, 1015–1036 (2010).

203. Perilli, M. *et al.* Novel TEM-type extended-spectrum beta-lactamase, TEM-134, in a Citrobacter koseri clinical isolate. *Antimicrob. Agents Chemother.* **49**, 1564–1566 (2005).

204. Bershtein, S., Goldin, K. & Tawfik, D. S. Intense neutral drifts yield robust and evolvable consensus proteins. *J. Mol. Biol.* **379**, 1029–1044 (2008).

205. Shcherbinin, D. S. *et al.* The study of the role of mutations M182T and Q39K in the TEM-72 β-lactamase structure by the molecular dynamics method. *Biochem. Moscow Suppl. Ser. B* **11**, 120–127 (2017).

206. San Millan, A. *et al.* Small-plasmid-mediated antibiotic resistance is enhanced by increases in plasmid copy number and bacterial fitness. *Antimicrob. Agents Chemother.* **59**, 3335–3341 (2015).

207. Haring, V. *et al.* Protein RepC is involved in copy number control of the broad host range plasmid RSF1010. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 6090–6094 (1985).

208. Jaffe, A., Chabbert, Y. A. & Semonin, O. Role of porin proteins OmpF and OmpC in the permeation of beta-lactams. *Antimicrob. Agents Chemother.* **22**, 942–948 (1982).

209. Couce, A., Rodríguez-Rojas, A. & Blázquez, J. Bypass of genetic constraints during mutator evolution to antibiotic resistance. *Proc. Biol. Sci.* **282**, 20142698 (2015).

210. Jaffé, A., Chabbert, Y. A. & Derlot, E. Selection and characterization of beta-lactam-resistant Escherichia coli K-12 mutants. *Antimicrob. Agents Chemother.* **23**, 622–625 (1983).

211. Ruiz, C. & Levy, S. B. Use of functional interactions with MarA to discover chromosomal genes affecting antibiotic susceptibility in Escherichia coli. *Int. J. Antimicrob. Agents* **37**, 177–178 (2011).

212. Hanoulle, X. *et al.* Structural analysis of Escherichia coli OpgG, a protein required for the biosynthesis of osmoregulated periplasmic glucans. *J. Mol. Biol.* **342**, 195–205 (2004).

213. Leying, H. J., Büscher, K. H., Cullmann, W. & Then, R. L. Lipopolysaccharide alterations responsible for combined quinolone and beta-lactam resistance in Pseudomonas aeruginosa. *Chemotherapy* **38**, 82–91 (1992).

214. Pagani, L., Landini, P., Luzzaro, F., Debiaggi, M. & Romero, E. Emergence of cross-resistance to imipenem and other beta-lactam antibiotics in Pseudomonas aeruginosa during therapy. *Microbiologica* **13**, 43–53 (1990).

215. Andrews, A. E., Lawley, B. & Pittard, A. J. Mutational analysis of repression and activation of the tyrP gene in Escherichia coli. *J. Bacteriol.* **173**, 5068–5078 (1991).

216. Wookey, P. J. & Pittard, A. J. DNA sequence of the gene (tyrP) encoding the tyrosine-specific transport system of Escherichia coli. *J. Bacteriol.* **170**, 4946–4949 (1988).

217. Brown, K. D. Formation of aromatic amino acid pools in Escherichia coli K-12. *J. Bacteriol.* **104**, 177–188 (1970).

218. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).

219. Weissborn, A. C. & Kennedy, E. P. Biosynthesis of membrane-derived oligosaccharides. Novel glucosyltransferase system from Escherichia coli for the elongation of beta 1----2-linked polyglucose chains. *J. Biol. Chem.* **259**, 12644–12651 (1984).

220. Braun, M. & Silhavy, T. J. Imp/OstA is required for cell envelope biogenesis in Escherichia coli. *Mol. Microbiol.* **45**, 1289–1302 (2002).

221. Bacher, J. M., Hughes, R. A., Tze-Fei Wong, J. & Ellington, A. D. Evolving new genetic codes. *Trends Ecol. Evol. (Amst.)* **19**, 69–75 (2004).

222. Rovner, A. J. *et al.* Recoded organisms engineered to depend on synthetic amino acids. *Nature* **518**, 89–93 (2015).

223. Mandell, D. J. *et al.* Biocontainment of genetically modified organisms by synthetic protein design. *Nature* **518**, 55–60 (2015).

224. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).

225. Yu, A. C.-S. *et al.* Mutations enabling displacement of tryptophan by 4-fluorotryptophan as a canonical amino acid of the genetic code. *Genome Biol Evol* **6**, 629–641 (2014).

226. Bacher, J. M. & Ellington, A. D. Selection and characterization of Escherichia coli variants capable of growth on an otherwise toxic tryptophan analogue. *J. Bacteriol.* **183**, 5414–5425 (2001).

227. Amiram, M. *et al.* Evolution of translation machinery in recoded bacteria enables multi-site incorporation of nonstandard amino acids. *Nat. Biotechnol.* **33**, 1272–1279 (2015).

228. Arai, K. *et al.* Preparation of Selenoinsulin as a Long-Lasting Insulin Analogue. *Angew. Chem. Int. Ed. Engl.* **56**, 5522–5526 (2017).

229. Muttenthaler, M. *et al.* Modulating oxytocin activity and plasma stability by disulfide bond engineering. *J. Med. Chem.* **53**, 8585–8596 (2010).

230. Metanis, N. & Hilvert, D. Strategic use of non-native diselenide bridges to steer oxidative protein folding. *Angew. Chem. Int. Ed. Engl.* **51**, 5585–5588 (2012).

231. Onderko, E. L., Silakov, A., Yosca, T. H. & Green, M. T. Characterization of a selenocysteine-ligated P450 compound I reveals direct link between electron donation and reactivity. *Nat Chem* **9**, 623–628 (2017).

232. Vandemeulebroucke, A., Aldag, C., Stiebritz, M. T., Reiher, M. & Hilvert, D. Kinetic consequences of introducing a proximal selenocysteine ligand into cytochrome P450cam. *Biochemistry* **54**, 6692–6703 (2015).

233. Hondal, R. J. Using chemical approaches to study selenoproteins-focus on thioredoxin reductases. *Biochim. Biophys. Acta* **1790**, 1501–1512 (2009).

234. Yu, Y. *et al.* Characterization and structural analysis of human selenium-dependent glutathione peroxidase 4 mutant expressed in Escherichia coli. *Free Radic. Biol. Med.* **71**, 332–338 (2014).

235. Mannes, A. M., Seiler, A., Bosello, V., Maiorino, M. & Conrad, M. Cysteine mutant of mammalian GPx4 rescues cell death induced by disruption of the wild-type selenoenzyme. *FASEB J.* **25**, 2135–2144 (2011).

236. Dery, L. *et al.* Accessing human selenoproteins through chemical protein synthesis †Electronic supplementary information (ESI) available. See DOI: 10.1039/c6sc04123j Click here for additional data file. *Chem Sci* **8**, 1922–1926 (2017).

237. Cohen, D. T., Zhang, C., Pentelute, B. L. & Buchwald, S. L. An Umpolung Approach for the Chemoselective Arylation of Selenocysteine in Unprotected Peptides. *J. Am. Chem. Soc.* **137**, 9784–9787 (2015).

238. Byrom, M., Bhadra, S., Jiang, Y. S. & Ellington, A. D. Exquisite allele discrimination by toehold hairpin primers. *Nucleic Acids Res.* **42**, e120 (2014).

239. Fürste, J. P. *et al.* Molecular cloning of the plasmid RP4 primase region in a multi-host-range tacP expression vector. *Gene* **48**, 119–131 (1986).

240. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

241. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol* **1151**, 165–188 (2014).

242. Salverda, M. L. M. *et al.* Initial mutations direct alternative pathways of protein evolution. *PLoS Genet.* **7**, e1001321 (2011).

243. Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Mol. Biol. Evol.* **33**, 268–280 (2016).

244. Blazquez, J., Morosini, M. I., Negri, M. C. & Baquero, F. Selection of naturally occurring extended-spectrum TEM beta-lactamase variants by fluctuating beta-lactam pressure. *Antimicrob. Agents Chemother.* **44**, 2182–2184 (2000).

245. Barlow, M. & Hall, B. G. Experimental prediction of the natural evolution of antibiotic resistance. *Genetics* **163**, 1237–1241 (2003).

246. Thai, Q. K., Bös, F. & Pleiss, J. The Lactamase Engineering Database: a critical survey of TEM sequences in public databases. *BMC Genomics* **10**, 390 (2009).

247. Forcella, C. *et al.* QnrB9 in association with TEM-116 extended-spectrum beta-lactamase in Citrobacter freundii isolated from sewage effluent: first report from Italy. *J Chemother* **22**, 243–245 (2010).

248. Corvec, S. *et al.* TEM-187, a new extended-spectrum β-lactamase with weak activity in a Proteus mirabilis clinical strain. *Antimicrob. Agents Chemother.* **57**, 2410–2412 (2013).

249. Paster, B. J., Olsen, I., Aas, J. A. & Dewhirst, F. E. The breadth of bacterial diversity in the human periodontal pocket and other oral sites. *Periodontol. 2000* **42**, 80–87 (2006).

250. Jenkinson, H. F. & Lamont, R. J. Oral microbial communities in sickness and in health. *Trends Microbiol.* **13**, 589–595 (2005).

251. Kolenbrander, P. E., Palmer, R. J., Periasamy, S. & Jakubovics, N. S. Oral multispecies biofilm development and the key role of cell-cell distance. *Nat. Rev. Microbiol.* **8**, 471–480 (2010).

252. Kolenbrander, P. E. *et al.* Communication among Oral Bacteria. *Microbiol Mol Biol Rev* **66**, 486–505 (2002).

253. Kolenbrander, P. E. Oral microbial communities: biofilms, interactions, and genetic systems. *Annu. Rev. Microbiol.* **54**, 413–437 (2000).

254. Huang, S. *et al.* Microbiota-based Signature of Gingivitis Treatments: A Randomized Study. *Sci Rep* **6**, (2016).

255. Moore, L. V. *et al.* Bacteriology of human gingivitis. *J. Dent. Res.* **66**, 989–995 (1987).

256. Huang, S. *et al.* Predictive modeling of gingivitis severity and susceptibility via oral microbiota. *ISME J* **8**, 1768–1780 (2014).

257. Kistler, J. O., Booth, V., Bradshaw, D. J. & Wade, W. G. Bacterial community development in experimental gingivitis. *PLoS ONE* **8**, e71227 (2013).

258. Sheiham, A. Is the chemical prevention of gingivitis necessary to prevent severe periodontitis? *Periodontol. 2000* **15**, 15–24 (1997).

259. Dumitrescu, A. L. Editorial: Periodontal Disease - A Public Health Problem. *Front Public Health* **3**, 278 (2015).

260. Abusleme, L. *et al.* The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. *ISME J* **7**, 1016–1025 (2013).

261. Ai, D. *et al.* Integrated metagenomic data analysis demonstrates that a loss of diversity in oral microbiota is associated with periodontitis. *BMC Genomics* **18**, (2017).

262. Dabdoub, S. M., Ganesan, S. M. & Kumar, P. S. Comparative metagenomics reveals taxonomically idiosyncratic yet functionally congruent communities in periodontitis. *Sci Rep* **6**, (2016).

263. Duran-Pinedo, A. E. *et al.* Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. *ISME J* **8**, 1659–1672 (2014).

264. Jorth, P. *et al.* Metatranscriptomics of the Human Oral Microbiome during Health and Disease. *mBio* **5**, (2014).

265. Li, Y. *et al.* Phylogenetic and functional gene structure shifts of the oral microbiomes in periodontitis patients. *ISME J* **8**, 1879–1891 (2014).

266. Shi, B. *et al.* Dynamic Changes in the Subgingival Microbiome and Their Potential for Diagnosis and Prognosis of Periodontitis. *mBio* **6**, (2015).

267. Szafranski, S. P. *et al.* High-Resolution Taxonomic Profiling of the Subgingival Microbiome for Biomarker Discovery and Periodontitis Diagnosis. *Appl Environ Microbiol* **81**, 1047–1058 (2015).

268. Wang, J. *et al.* Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. *Sci Rep* **3**, 1843 (2013).

269. Yost, S., Duran-Pinedo, A. E., Teles, R., Krishnan, K. & Frias-Lopez, J. Functional signatures of oral dysbiosis during periodontitis progression revealed by microbial metatranscriptome analysis. *Genome Med* **7**, 27 (2015).

270. Huang, S. *et al.* Preliminary characterization of the oral microbiota of Chinese adults with and without gingivitis. *BMC Oral Health* **11**, 33 (2011).

271. Meuric, V. *et al.* Signature of Microbial Dysbiosis in Periodontitis. *Appl Environ Microbiol* **83**, (2017).

272. Kim, S.-H. *et al.* Improved accuracy in periodontal pocket depth measurement using optical coherence tomography. *J Periodontal Implant Sci* **47**, 13–19 (2017).

273. Blazewicz, S. J., Barnard, R. L., Daly, R. A. & Firestone, M. K. Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. *ISME J* **7**, 2061–2068 (2013).

274. Szafrański, S. P. *et al.* Functional biomarkers for chronic periodontitis and insights into the roles of Prevotella nigrescens and Fusobacterium nucleatum; a metatranscriptome analysis. *NPJ Biofilms Microbiomes* **1**, 15017 (2015).

275. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

276. Wassenaar, T. M. & Gaastra, W. Bacterial virulence: can we draw the line? *FEMS Microbiol. Lett.* **201**, 1–7 (2001).

277. Dumitrescu, A. L. Etiology and pathogenesis of periodontal disease. *Springer Science & Business Media* 39–76 (2009).

278. Smith, I. Mycobacterium tuberculosis Pathogenesis and Molecular Determinants of Virulence. *Clin Microbiol Rev* **16**, 463–496 (2003).

279. Kajfasz, J. K. *et al.* Two Spx Proteins Modulate Stress Tolerance, Survival, and Virulence in Streptococcus mutans. *J Bacteriol* **192**, 2546–2556 (2010).

280. Singer, R. E. & Buckner, B. A. Butyrate and propionate: important components of toxic dental plaque extracts. *Infect Immun* **32**, 458–463 (1981).

281. Górska, R. *et al.* Relationship between clinical parameters and cytokine profiles in inflamed gingival tissue and serum samples from patients with chronic periodontitis. *J. Clin. Periodontol.* **30**, 1046–1052 (2003).

282. Sculley, D. V. & Langley-Evans, S. C. Periodontal disease is associated with lower antioxidant capacity in whole saliva and evidence of increased protein oxidation. *Clin. Sci.* **105**, 167–172 (2003).

283. Leggott, P. J., Robertson, P. B., Rothman, D. L., Murray, P. A. & Jacob, R. A. The effect of controlled ascorbic acid depletion and supplementation on periodontal health. *J. Periodontol.* **57**, 480–485 (1986).

284. Tancharoen, S. *et al.* Cleavage of host cytokeratin-6 by lysine-specific gingipain induces gingival inflammation in periodontitis patients. *PLoS ONE* **10**, e0117775 (2015).

285. de Diego, I. *et al.* Structure and Mechanism of Cysteine Peptidase Gingipain K (Kgp), a Major Virulence Factor of Porphyromonas gingivalis in Periodontitis. *J Biol Chem* **289**, 32291–32302 (2014).

286. Bengtsson, T., Khalaf, A. & Khalaf, H. Secreted gingipains from Porphyromonas gingivalis colonies exert potent immunomodulatory effects on human gingival fibroblasts. *Microbiol. Res.* **178**, 18–26 (2015).

287. Paige, M. *et al.* Role of leukotriene A4 hydrolase aminopeptidase in the pathogenesis of emphysema. *J Immunol* **192**, 5059–5068 (2014).

288. Sakihama, Y., Mizoguchi, H., Oshima, T. & Ogasawara, N. YdfH identified as a repressor of rspA by the use of reduced genome Escherichia coli MGF-01. *Biosci. Biotechnol. Biochem.* **76**, 1688–1693 (2012).

289. Socransky, S. S. & Haffajee, A. D. Periodontal microbial ecology. *Periodontol. 2000* **38**, 135–187 (2005).

290. Socransky, S. S., Haffajee, A. D., Cugini, M. A., Smith, C. & Kent, R. L. Microbial complexes in subgingival plaque. *J. Clin. Periodontol.* **25**, 134–144 (1998).

291. Couturier, M. R., Slechta, E. S., Goulston, C., Fisher, M. A. & Hanson, K. E. Leptotrichia Bacteremia in Patients Receiving High-Dose Chemotherapy. *J Clin Microbiol* **50**, 1228–1232 (2012).

292. Eribe, E. R. K. & Olsen, I. Leptotrichia species in human infections II. *J Oral Microbiol* **9**, (2017).

293. Valour, F. *et al.* Actinomycosis: etiology, clinical features, diagnosis, treatment, and management. *Infect Drug Resist* **7**, 183–197 (2014).

294. Smalley, J. W., Birss, A. J., Szmigielski, B. & Potempa, J. Sequential action of R- and K-specific gingipains of Porphyromonas gingivalis in the generation of the haem-containing pigment from oxyhaemoglobin. *Arch. Biochem. Biophys.* **465**, 44–49 (2007).

295. Stathopoulou, P. G., Benakanakere, M. R., Galicia, J. C. & Kinane, D. F. Epithelial cell pro-inflammatory cytokine response differs across dental plaque bacterial species. *J. Clin. Periodontol.* **37**, 24–29 (2010).

296. Oetjen, J., Fives-Taylor, P. & Froeliger, E. Characterization of a Streptococcal Endopeptidase with Homology to Human Endothelin-Converting Enzyme. *Infect Immun* **69**, 58–64 (2001).

297. Holmlund, A., Lampa, E. & Lind, L. Poor Response to Periodontal Treatment May Predict Future Cardiovascular Disease. *J. Dent. Res.* **96**, 768–773 (2017).

298. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).

299. Dodt, M., Roehr, J. T., Ahmed, R. & Dieterich, C. FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology (Basel)* **1**, 895–905 (2012).

300. Chen, T. *et al.* The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)* **2010**, baq013 (2010).

301. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**, 29–34 (1999).

302. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

303. Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A. & Kocher, J.-P. Calculating sample size estimates for RNA sequencing data. *J. Comput. Biol.* **20**, 970–978 (2013).

304. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).